# Active Learning and Basis Selection for Kernel-Based Linear Models: A Bayesian Perspective

John Paisley, Xuejun Liao and Lawrence Carin
Department of Electrical and Computer Engineering
Duke University, Durham, NC
{jwp4,xjliao,lcarin}@ece.duke.edu

## Abstract

We develop an active learning algorithm for kernel-based linear regression and classification. The proposed greedy algorithm employs a minimum-entropy criterion derived using a Bayesian interpretation of ridge regression. We assume access to a matrix, $\mathbf{\Phi} \in \mathbb{R}^{N \times N}$, for which the $(i,j)^{th}$ element is defined by the kernel function $K(\gamma_i, \gamma_j) \in \mathbb{R}$, with the observed data $\gamma_i \in \mathbb{R}^d$. We seek a model, $\mathcal{M} : \gamma_i \to y_i$, where $y_i$ is a real-valued response or integer-valued label, which we do not have access to *a priori*. To achieve this goal, a sub-matrix, $\mathbf{\Phi}_{I_l, I_b} \in \mathbb{R}^{n \times m}$, is sought that corresponds to the intersection of $n$ rows and $m$ columns of $\mathbf{\Phi}$, indexed by the sets $I_l$ and $I_b$ respectively. Typically $m \ll N$ and $n \ll N$.

We have two objectives: $(i)$ Determine the $m$ columns of $\mathbf{\Phi}$, indexed by the set $I_b$, that are the most informative for building a linear model, $\mathcal{M} : [1 \ \mathbf{\Phi}_{i,I_b}]^T \to y_i$, without any knowledge of $\{y_i\}_{i=1}^N$ and $(ii)$ using active learning, sequentially determine which subset of $n$ elements of $\{y_i\}_{i=1}^N$ should be acquired; both stopping values, $|I_b| = m$ and $|I_l| = n$, are also to be inferred from the data. These steps are taken with the goal of minimizing the uncertainty of the model parameters, $x$, as measured by the differential entropy of its posterior distribution. The parameter vector $x \in \mathbb{R}^m$, as well as the model bias $\eta \in \mathbb{R}$, is then learned from the resulting problem, $y_{I_l} = \mathbf{\Phi}_{I_l, I_b} x + \eta \mathbf{1} + \epsilon$. The remaining $N - n$ responses/labels not included in $y_{I_l}$ can be inferred by applying $x$ to the remaining $N - n$ rows of $\mathbf{\Phi}_{:, I_b}$. We show experimental results for several regression and classification problems, and compare with other active learning methods.

## Index Terms

active learning, linear regression & classification, Bayesian models, kernel methods, optimal experiments

## I. Introduction

Linear models provide a popular framework for performing regression and classification [20]. Given a completely specified (e.g., labeled) data set, $\mathcal{D}_c = \{(\gamma_i, y_i)\}_{i=1}^N$, where $\gamma_i \in \mathbb{R}^d$ and $y_i$ is a real-valued response, or binary label associated with $\gamma_i$, these models take the form,

$$y = \mathbf{\Phi} x + \eta \mathbf{1} + \epsilon \tag{1}$$

where $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$, and $\eta \in \mathbb{R}$ accounts for the model bias. For classification using a probit function [20], the output of the linear model is a vector of latent variables, where each entry determines the corresponding class membership by its value with respect to a threshold. This does not change the analysis of this paper, and so, for simplicity, we allow $y$ in (1) to take the values of the binary labels. If $\mathbf{\Phi}$ is constituted using the raw data, with the $i^{th}$ row of $\mathbf{\Phi}$ corresponding to $\gamma_i$, then $M = d$. For nonlinear decision boundaries in $\mathbb{R}^d$, a kernel function can be used, $K(\gamma_i, \gamma_j)$, such as the Gaussian kernel,

$$K(\gamma_i, \gamma_j) = \exp\left[\frac{-\|\gamma_i - \gamma_j\|_2^2}{\upsilon^2}\right] \tag{2}$$

in which case $\mathbf{\Phi}(i, j) = K(\gamma_i, \gamma_j)$. Typically, the number of samples, $N$, is much larger than the feature dimensionality, $d$, with the dimensionality of the model parameters, $x$, increased accordingly.

To avoid over-fitting the model to the data, it is often imposed that $x$ should be *sparse*, or that most of the elements of $x$ be set equal to zero. To this end, sparse models such as kernel matching pursuits (KMP) [27], the support vector machine (SVM) [4], the relevance vector machine (RVM) [25] and the LASSO [24] have been introduced for regression and classification. These models infer a subset of $m$ points within the data set on which to build a kernel. This process reduces $\mathbf{\Phi}$ to the smaller matrix, $\mathbf{\Phi}_{:,I_b} \in \mathbb{R}^{N \times m}$, where $I_b$ is a set of integers that index which columns, or basis functions, of $\mathbf{\Phi}$ are selected, and the symbol ':' denotes that all rows are selected. This reduces the problem to

$$y = \mathbf{\Phi}_{:,I_b} x + \eta \mathbf{1} + \epsilon \tag{3}$$

where $|I_b| = m \ll N$ is a number learned in the model building process. As mentioned, a benefit of this sparseness is a better generalization to new data, which is achieved by choosing from $\mathcal{D} = \{\gamma_i\}_{i=1}^N$ a subset of the $m$ most relevant points for characterizing the data set.

A potential drawback of these methods is that they require access to the corresponding vector of labels/responses, $\{y_i\}_{i=1}^N$, which may be unavailable and difficult to obtain. For example, in a medical diagnostic problem, we may posses a large data set, $\mathcal{D}$, containing the symptoms of individual patients, without possessing the corresponding truth, $\{y_i\}_{i=1}^N$, regarding their medical condition. When we do not have these values, the algorithms for designing sparse models mentioned above are no longer feasible.

The challenge of inferring which elements of $\{y_i\}_{i=1}^N$ to acquire, with the goal of maximally reducing model uncertainty, is called *active learning* [6]. However, in most active learning research to date, it has been assumed that the model parameters of interest are known in advance, with measurement locations selected to most efficiently reduce the uncertainty (e.g., entropy) of these parameters. For linear models, these parameters correspond to the coefficient vector, $x$, which is usually multiplied directly with the data matrix [18]. However, for learning linear models using a kernel, we must first define the index set, $I_b$, of coefficients in $x$ that we are interested in and construct the matrix $\mathbf{\Phi}_{:,I_b}$, before measurements can be made. Thus a circular dependence arises.

In this paper, we discuss both aspects of this problem. We term the construction of $I_b$ *basis selection*, though we view this process as part of a larger active learning problem. Assuming no *a priori* values for $\{y_i\}_{i=1}^N$, we $(i)$ define the matrix $\mathbf{\Phi}_{:,I_b}$, which in effect defines the $x$ we wish to learn and $(ii)$ define which $y_i$ should be acquired to most effectively reduce the uncertainty in the values of $x$. Defining the set $I_l$ to contain the indices of the data for which we obtain measurements, then the vector $y_{I_l} \in \mathbb{R}^n$ contains these $|I_l| = n \ll N$ values acquired through the active learning process, and $\mathbf{\Phi}_{I_l,I_b} \in \mathbb{R}^{n \times m}$ represents the corresponding matrix. Our goal using active learning, including basis selection, is to find this underlying linear model that most efficiently represents and solves the problem,

$$y_{I_l} = \mathbf{\Phi}_{I_l,I_b} x + \eta \mathbf{1} + \epsilon \tag{4}$$

with $x$ solved using regularized least squares, e.g., ridge regression, [3], [13], among other potential solutions.

We focus on the ridge regression solution since it provides a framework for defining an efficient, greedy procedure for determining the matrix, $\mathbf{\Phi}_{I_l,I_b}$, as well as the corresponding measurement locations, $y_{I_l}$, on which to build a regression or classification model. This is done according to an entropy measure that naturally arises from the Bayesian interpretation of these two solutions. This measure results in a process whereby the relevant vectors are selected (defining the set $I_b$) followed by the measurement locations (defining the set $I_l$). This paper is a continuation of previous work [29], [30], [17], [16], with the intended contribution being an analysis of the approach in Bayesian terms, including the connection of active measurement selection with the Gaussian process, leading to further analysis of measurement selection, along with a more developed Bayesian model hierarchy and an extension to the regression problem.

The remainder of the paper is organized as follows. In Section 2, we review the least squares and ridge regression solutions to overdetermined linear systems. This allows us to state the active learning problem in probabilistic terms and emphasize the Bayesian aspect of the problem. In Section 3, after defining our optimization criterion, we present an active learning algorithm for kernel-based linear regression and classification, which entails iterative methods for basis selection and measurement selection. This is followed in Section 4 by analysis of a synthesized data set, and applications to real data for regression and classification. We conclude the paper in Section 5.

## II. Least Squares and Ridge Regression with Bayesian Interpretations

In this section, we review least squares and ridge regression for linear models and discuss their Bayesian interpretations. This will provide the theoretical foundation for the active learning algorithm of the following section. For clarity, we assume that $\bar{y} = 0$, and therefore $\eta = 0$. Though this assumption is not reasonable in the context of active learning, it does not affect our analysis, and we will account for $\eta$ in later sections. Consider a full rank, $N \times m$ matrix, $\mathbf{\Phi}$, with $m \leq N$, and the overdetermined linear system,

$$y = \mathbf{\Phi} x + \epsilon \tag{5}$$

where $y \in \mathbb{R}^N$, $x \in \mathbb{R}^m$ and $\epsilon$ is an $N$-dimensional error vector. Typically, $y$ does not reside within the subspace spanned by the columns of $\mathbf{\Phi}$, and so an approximation to $y$ is desired, $\hat{y} = \mathbf{\Phi} x$, using an $x$ that meets certain predefined criteria. Least squares [3] and ridge regression [13] are two particular solutions, which we review below.

### A. *Least Squares*

Least squares is a well-known approximate solution for $y = \mathbf{\Phi} x + \epsilon$ that finds the value of $x$ which minimizes the total squared error in approximating $y$,

$$x_{\text{LS}} = \arg\min_{x} \|y - \mathbf{\Phi} x\|_2^2 \tag{6}$$

Calling this objective function $f_{\text{LS}}(x)$, this solution can be found by setting $\nabla_x f_{\text{LS}}(x) = 0$, or by using the vector space interpretation, where the error vector, $\epsilon = y - \mathbf{\Phi} x$, is orthogonal to the approximation, $\mathbf{\Phi} x$, which results in a dot product that is equal to zero.

The least squares solution can lead to undesirable properties of $x$ [13], which can be seen by interpreting $y = \mathbf{\Phi} x + \epsilon$ as a generative process. Under this interpretation, the vector $\epsilon$ is modeled as a Gaussian noise vector, $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$, allowing the linear system to be rewritten as,

$$y \sim \mathcal{N}\left(\mathbf{\Phi} x, \sigma^2 I\right) \tag{7}$$

When $x$ is of interest, rather than $y$, this interpretation of (5) can produce $x_{\text{LS}}$ approximations that deviate significantly from the underlying truth. For example, under the interpretation of (7), the expectation and covariance of $x_{\text{LS}}$ is,

$$\mathbb{E}[x_{\text{LS}}] = x, \qquad \mathbb{V}[x_{\text{LS}}] = \sigma^2 \left(\mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \tag{8}$$

When $\mathbf{\Phi}^T \mathbf{\Phi}$ has eigenvalues that are very small, this can lead to extremely large values in the covariance matrix. The approximation to $y$ might be good, but an $x_{\text{LS}}$ might be used that is far from the true $x$ in a Euclidean sense. This problem can be resolved if additional error is allowed in the approximation to $y$.

### *B. Ridge Regression*

To address this issue, an $\ell_2$ regularized least squares solution, called ridge regression or Tikhonov regularization, can be used [13][3]. It alters the least squares objective function of (6) by adding a penalty term for the magnitude of $x$. In this formulation, the solution, $x_{\text{RR}}$, can be found by minimizing

$$x_{\text{RR}} = \arg\min_{x} \; \|y - \mathbf{\Phi} x\|_2^2 + \beta x^T x \tag{9}$$

where the first term is the least squares objective function and $\beta$ is a positive scalar that penalizes increasing magnitudes of $x$. As $\beta \rightarrow 0$, we see that $x_{\text{RR}} \rightarrow x_{\text{LS}}$. Calling this function $f_{RR}(x)$, we can calculate $x_{\text{RR}}$ analytically by setting $\nabla_x f_{RR}(x) = 0$. The result is,

$$x_{\text{RR}} = \left(\beta I + \mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T y \tag{10}$$

The value, $\beta$, can be viewed as a lower bound on the eigenvalues of $\left(\beta I + \mathbf{\Phi}^T \mathbf{\Phi}\right)$ because the eigenvalues of this matrix are equal to the sum of $\beta$ and the eigenvalues of $\mathbf{\Phi}^T \mathbf{\Phi}$. As a result, the inverse is more constrained in the magnitude of it's values and $x_{\text{RR}}$ does not “blow up” as easily. Sacrificed in this process is the unbiased nature of the least squares estimate, since $\mathbb{E}[x_{\text{RR}}] \neq x$.

### *C. ML, MAP and Bayesian Interpretations of Least Squares and Ridge Regression*

Least squares and ridge regression can be motivated as maximum likelihood and maximum *a posteriori* solutions, respectively, which are then naturally extended to the fully Bayesian setting [2]. First, recall the generative interpretation,

$$y \sim \mathcal{N}\left(\mathbf{\Phi} x, \sigma^2 I\right) \tag{11}$$

where we here assume that $y, \mathbf{\Phi}$, and $\sigma^2$ are known. This normal distribution serves as a likelihood function for possible $x$ values. The maximum likelihood (ML) estimate of $x$, written as

$$\nabla_x \ln p\left(y|x, \mathbf{\Phi}, \sigma^2\right)|_{x_{\text{ML}}} = 0 \tag{12}$$

is equivalent to maximizing the negative of the least squares penalty function, and thus $x_{\text{ML}} = x_{\text{LS}}$.

In the Bayesian setting, one can place a prior distribution, $p(x)$, on $x$. Given that the likelihood in (11) is normal, a conjugate, zero-mean normal prior distribution can be used,

$$x \sim \mathcal{N}\left(0, \alpha^{-1} I\right) \tag{13}$$

with the maximum a posteriori (MAP) estimate then found by solving

$$\nabla_x \left[\ln p\left(y|x, \mathbf{\Phi}, \sigma^2\right) + \ln p(x|\alpha)\right] |_{x_{\text{MAP}}} = 0 \tag{14}$$

This is equivalent to maximizing the negative of the ridge regression penalty function with $\beta \equiv \alpha\sigma^2$. Therefore, $x_{\text{MAP}} = x_{\text{RR}}$ when a zero-mean, multivariate normal prior is placed on $x$.

The Bayesian interpretation arises by considering the posterior distribution of $x$, which can be calculated analytically according to Bayes' rule. The posterior distribution of $x$ is proportional to the product of the likelihood and the prior,

$$p\left(x|y, \mathbf{\Phi}, \alpha, \sigma^2\right) \propto p\left(y|x, \mathbf{\Phi}, \sigma^2\right) p(x|\alpha) \tag{15}$$

Due to conjugacy, the posterior distribution of $x$ is multivariate normal with mean, $\mu$, and covariance, $\Sigma$, equal to

$$\mu = \left(\alpha\sigma^2 I + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T y \tag{16}$$

$$\Sigma = \left(\alpha I + \frac{1}{\sigma^2}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1} \tag{17}$$

We see that the ridge regression, or MAP solution is the expectation of the posterior distribution of $x$ in this Bayesian formulation. We will refer to this full posterior as the "Bayesian ridge" solution. In the next section, we show how this framework provides a natural means for performing both basis selection and measurement selection that is optimal according to a minimum-entropy criterion.

## III. Optimal Active Learning for Kernel-Based Linear Regression and Classification

Restating the problem, we are given a data set of $N$ observations, $\mathcal{D} = \{\gamma_i\}_{i=1}^N$, where $\gamma_i \in \mathbb{R}^d$, and we define a kernel function, $K(\gamma_i, \gamma_j)$, between points $\gamma_i$ and $\gamma_j$. These values are contained in the matrix $\mathbf{\Phi} \in \mathbb{R}^{N\times N}$. We also assume a corresponding set of *missing* labels or responses, $\{y_i\}_{i=1}^N$, one for each observation $\gamma_i$, with the complete data set being $\mathcal{D}_c = \{(\gamma_i, y_i)\}_{i=1}^N$. We are interested in constructing a matrix, $\mathbf{\Phi}_{I_l,I_b} \in \mathbb{R}^{n\times m}$, consisting of the intersection of $n$ rows and $m$ columns of the complete matrix, $\mathbf{\Phi}$, indexed by the sets $I_l$ and $I_b$ respectively. We also include a vector, $\eta\mathbf{1}$, for the model bias and note that the addition of this vector alters equation (16) by replacing $y$ with $y - \eta\mathbf{1}$ and leaves (17) unchanged.

The two aspects of our active learning approach, basis selection (or defining the set $I_b$) and measurement selection (or defining the set $I_l$), are intended to characterize $x$ as efficiently as possible. We define our measure of efficiency via a quantification of the information gained at each step using the differential entropy of $x$.

### A. *Differential Entropy as a Measure of Information Gain*

The differential entropy of a continuous random variable is a measure of uncertainty in its value [7] and is a popular measure of information gain for active learning [18], [15]. The smaller, or more negative this value is, the less the uncertainty in the volume of the space in which the corresponding random variable can live. For a Gaussian random variable in $\mathbb{R}^m$ having the covariance matrix $\Sigma$, the differential entropy is equal to,

$$h(x) = \frac{1}{2} \ln \left[ (2\pi e)^m |\Sigma| \right] \tag{18}$$

With the addition of each basis function or response/label acquisition to the model, the change in differential entropy measures the impact of that particular basis or response/label on our uncertainty in $x$. When viewing the linear model from the Bayesian perspective outlined in Section II-C, the differential entropy function naturally arises as a measure of information gain. We observe that only the posterior covariance matrix of $x$ factors into this measure, which has the important property of being independent of both $\{y_i\}_{i=1}^N$ and the model bias, $\eta$. In the following sections, we discuss basis selection and measurement selection using this differential entropy measure.

### B. *Optimal Basis Selection Using a Greedy Selection Criterion*

With optimal basis selection, we seek a subset of columns, $\mathbf{\Phi}_{:,I_b}$, where $I_b \subset \{1, \ldots, N\}$, in a way that best characterizes the space of the data set. These columns effectively define the subset of $\mathcal{D}$ to be used as kernel functions. After $m$ steps of this process, the matrix $\mathbf{\Phi}_{:,I_b} \in \mathbb{R}^{N \times m}$ is the most efficient $m$-basis representation of $y = \mathbf{\Phi}_{:,I_b} x + \eta \mathbf{1} + \epsilon$ under our myopic selection criterion. Stated another way, we are interested in finding the $m$ most relevant vectors when $y$ is unknown [25].

As previously mentioned, two observations play an important role in basis selection: ($i$) The posterior differential entropy of $x$ under the Bayesian ridge interpretation only depends on the posterior covariance matrix, $\Sigma$, and ($ii$) this posterior covariance matrix does not depend on the label/response vector, $y$, or the model bias, $\eta$. Therefore, in considering the differential entropy, $h_k(x)$, at step $k$, only the resulting covariance matrix,

$$\Sigma_k = \left( \alpha I + \sigma^{-2} \mathbf{\Phi}^T_{:,I_b^{(k)}} \mathbf{\Phi}_{:,I_b^{(k)}} \right)^{-1}$$

needs to be considered when quantifying the information gain. In this and the following section, we use the notation $I_b^{(k)}$ to indicate that $|I_b| = k$, and $I_b(k)$ to select the $k^{th}$ element of $I_b$.

Given a set of $k$ vectors, $\mathbf{\Phi}_{:,I_b^{(k)}}$, it follows that we would like to select the $(k+1)^{st}$ column vector from $\mathbf{\Phi}$ that minimizes the uncertainty of $x$ at step $k+1$, as quantified by the differential entropy, $h_{k+1}(x)$. Equation (18) indicates that this is done by selecting this vector with the goal of minimizing the determinant of the posterior covariance matrix,

$$|\Sigma_{k+1}| = \left| \left( \alpha I + \sigma^{-2} \mathbf{\Phi}^T_{:,I_b^{(k+1)}} \mathbf{\Phi}_{:,I_b^{(k+1)}} \right)^{-1} \right|$$

For basis selection, we let $\alpha \to 0$, which corresponds to a noninformative prior on $x$, allowing the data to completely inform the selection of basis functions. Therefore, the next basis function can equivalently be selected to maximize $\left| \mathbf{\Phi}^T_{:,I_b^{(k+1)}} \mathbf{\Phi}_{:,I_b^{(k+1)}} \right|$. We separate $\mathbf{\Phi}_{:,I_b^{(k+1)}}$ into the currently active set, $\mathbf{\Phi}_{:,I_b^{(k)}}$, and the proposed basis function, $\phi_i$, where $i$ indexes a column of $\mathbf{\Phi}$ and is the proposed $(k+1)^{st}$ index element in $I_b^{(k+1)}$. This matrix multiplication can then be written as,

$$\mathbf{\Phi}^T_{:,I_b^{(k+1)}} \mathbf{\Phi}_{:,I_b^{(k+1)}} = \begin{bmatrix} \mathbf{\Phi}^T_{:,I_b^{(k)}} \mathbf{\Phi}_{:,I_b^{(k)}} & \mathbf{\Phi}^T_{:,I_b^{(k)}} \phi_i \\ \phi_i^T \mathbf{\Phi}_{:,I_b^{(k)}} & \phi_i^T \phi_i \end{bmatrix} \tag{19}$$

The determinant of this right matrix can be expressed using the following identity,

$$\begin{vmatrix} \mathbf{\Phi}^T_{:,I_b^{(k)}} \mathbf{\Phi}_{:,I_b^{(k)}} & \mathbf{\Phi}^T_{:,I_b^{(k)}} \phi_i \\ \phi_i^T \mathbf{\Phi}_{:,I_b^{(k)}} & \phi_i^T \phi_i \end{vmatrix} = \left| \mathbf{\Phi}^T_{:,I_b^{(k)}} \mathbf{\Phi}_{:,I_b^{(k)}} \right| \left( \phi_i^T \phi_i - \phi_i^T \mathbf{\Phi}_{:,I_b^{(k)}} \left( \mathbf{\Phi}^T_{:,I_b^{(k)}} \mathbf{\Phi}_{:,I_b^{(k)}} \right)^{-1} \mathbf{\Phi}^T_{:,I_b^{(k)}} \phi_i \right) \tag{20}$$

To maximize this term for a new vector $\phi_i$, selected from $\mathbf{\Phi}$, we therefore must maximize the term in parenthesis on the right side of (20), leading to the following definition.

***Optimal Basis Selection***: Given a matrix of selected basis functions, or column vectors, $\mathbf{\Phi}_{:,I_b^{(k)}}$, indexed by the set $I_b^{(k)}$, the $(k+1)^{st}$ index value, $I_b(k+1)$, is selected according to the following function of the columns, $\phi_i$, of $\mathbf{\Phi}$,

$$I_b(k+1) = \underset{i \in \{1,\ldots,N\}}{\arg\max} \ \phi_i^T \phi_i - \phi_i^T \mathbf{\Phi}_{:,I_b^{(k)}} \left( \mathbf{\Phi}^T_{:,I_b^{(k)}} \mathbf{\Phi}_{:,I_b^{(k)}} \right)^{-1} \mathbf{\Phi}^T_{:,I_b^{(k)}} \phi_i \tag{21}$$

which is the basis that minimizes the differential entropy of the posterior distribution of $x$ according to the Bayesian ridge approach.

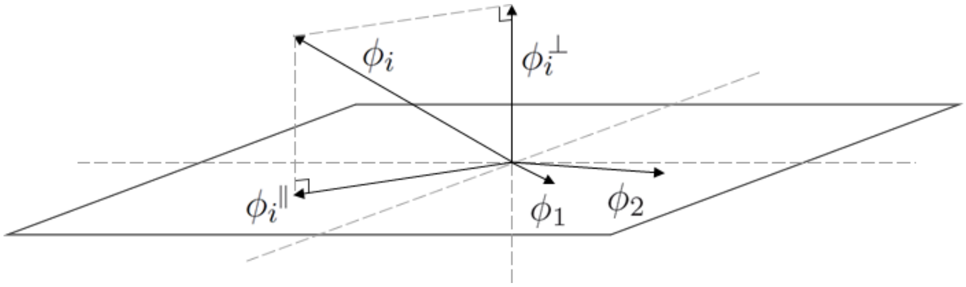


Fig. 1. Optimal basis selection at step $k+1$: Select the basis having the largest component, $\|\phi_i^\perp\|$, perpendicular to the hyperplane spanned by $\mathbf{\Phi}_{:,I_b^{(k)}}$, being the set of basis vectors selected through step $k$. In this illustration, $k = 2$.

This new basis function has an interesting vector space interpretation, which is illustrated in Figure 1. In (21), the term $\phi_i^T \mathbf{\Phi}_{:,I_b^{(k)}} \left( \mathbf{\Phi}_{:,I_b^{(k)}}^T \mathbf{\Phi}_{:,I_b^{(k)}} \right)^{-1} \mathbf{\Phi}_{:,I_b^{(k)}}^T \phi_i$ corresponds to the squared magnitude of $\phi_i^\|$, the component of $\phi_i$ projected onto the hyperplane spanned by $\mathbf{\Phi}_{:,I_b^{(k)}}$. The term $\phi_i^T \phi_i$ is simply the squared magnitude of $\phi_i$. Since these vectors form a right triangle, the difference of these two values, $\|\phi_i^\perp\|_2^2$, is the squared magnitude of the component of $\phi_i$ perpendicular to the subspace spanned by the vectors currently selected to represent the linear system, $\mathbf{\Phi}_{:,I_b^{(k)}}$. The optimal next basis therefore corresponds to the column vector, $\phi_i$, contained in $\mathbf{\Phi}$, with the largest component projected into the null space of $\mathbf{\Phi}_{:,I_b^{(k)}}$. As this value will be zero for any column indexed by $I_b^{(k)}$, we see that a unique vector will be selected with each iteration. We therefore do not have to worry about repeated index values in $I_b$. Another way to state this is that $\|\phi_i^\perp\|_2^2$ is equal to the squared magnitude of the error vector that results from a least squares approximation of $\phi_i$ using $\mathbf{\Phi}_{:,I_b^{(k)}}$.

We contrast this basis selection process with matching pursuits algorithms [26],[27], which require the information contained in $y$. In matching pursuits, the selected basis at iteration $k+1$ is that which is most parallel to the error vector, $y - \mathbf{\Phi}_{:,I_b^{(k)}} x_{\text{LS}}^{(k)}$, where $x_{\text{LS}}^{(k)}$ is the least squares solution computed after step $k$. Due to the orthogonality of the error, the ideal next vector will have the property of being nearly orthogonal to $\mathbf{\Phi}_{:,I_b^{(k)}}$. In active basis selection (21), the orthogonality of $\phi_{I_b(k+1)}$ to $\mathbf{\Phi}_{:,I_b^{(k)}}$ is not an important factor, but rather how much it extends into the null space of $\mathbf{\Phi}_{:,I_b^{(k)}}$. This is intuitively reasonable, since there is a potential tradeoff between the orthogonality of $\phi_{I_b(k+1)}$ and the amount of information $\phi_{I_b(k+1)}$ contains about the values of $y$ for neighboring points in the kernel. For example, if $\gamma^* \in \mathcal{D}$ is an outlier, its basis function, $\phi^*$, may be nearly orthogonal to $\mathbf{\Phi}_{:,I_b^{(k)}}$, but contain virtually no information about neighboring labels/responses, as indicated by its small magnitude. This point will therefore not be selected.

We discuss two possible stopping criteria for basis selection. The first is the change in differential entropy, $h_{k+1}(x) - h_k(x)$, which is a measure of the information gained by adding the $(k+1)^{st}$ basis function. Again letting $\alpha \rightarrow 0$, this is equal to,

$$h_{k+1}(x) - h_k(x) = \frac{1}{2}\ln(2\pi e\sigma^2) - \frac{1}{2}\ln\left(\|\phi^{\perp}_{I_b(k+1)}\|_2^2\right) \tag{22}$$

where we use the definition of $\|\phi^{\perp}_{I_b(k+1)}\|_2^2$ as the squared magnitude of the projection of the $(k+1)^{st}$ basis onto the null space of $\mathbf{\Phi}_{:,I_b^{(k)}}$. Basis selection can be terminated when $h_{k+1}(x) - h_k(x) > 0$. Using a Gaussian kernel, we can also choose to terminate when $\frac{1}{2}\ln\left(\|\phi^{\perp}_{I_b(k+1)}\|_2^2\right) < 0$, which occurs when $\|\phi^{\perp}_{I_b(k+1)}\|_2^2 < 1$. This will ensure that all data is represented by basis function; since the diagonal of $\mathbf{K}$ is equal to one, any observation that is not represented will have a null space magnitude of at least one. Of course, the values along each row can also be monitored directly.

Another possible termination criterion is the inverse condition number of the matrix, $\mathbf{\Phi}^T_{:,I_b^{(k+1)}}\mathbf{\Phi}_{:,I_b^{(k+1)}}$,

$$\kappa^{-1}\left(\mathbf{\Phi}^T_{:,I_b^{(k+1)}}\mathbf{\Phi}_{:,I_b^{(k+1)}}\right) = \frac{\lambda_{\text{min}}\left(\mathbf{\Phi}^T_{:,I_b^{(k+1)}}\mathbf{\Phi}_{:,I_b^{(k+1)}}\right)}{\lambda_{\text{max}}\left(\mathbf{\Phi}^T_{:,I_b^{(k+1)}}\mathbf{\Phi}_{:,I_b^{(k+1)}}\right)} \tag{23}$$

where the $\lambda\left(\mathbf{\Phi}^T_{:,I_b^{(k+1)}}\mathbf{\Phi}_{:,I_b^{(k+1)}}\right)$ values denote the largest and smallest eigenvalues of $\mathbf{\Phi}^T_{:,I_b^{(k+1)}}\mathbf{\Phi}_{:,I_b^{(k+1)}}$, or the square of the largest and smallest singular values of $\mathbf{\Phi}_{:,I_b^{(k+1)}}$. The condition number of a matrix is a measure of how easily $x$ can be inferred given the output $y = \mathbf{\Phi}_{:,I_b^{(k+1)}}x + \eta\mathbf{1} + \epsilon$. The final index set, $I_b^{(m)}$ represents the set of points in $\mathcal{D}$ to be stored for future basis function calculations for new data. If the data set, $\mathcal{D}$, is representative of the entire space of interest, we can be confident that the selected basis functions will be able to represent all future observations.

### *C. Optimal Measurement Selection Using a Greedy Selection Criterion*

Following the basis selection process of Section III-B, which defines the index set $I_b$, the next task is to select points, $\gamma_i$, from the data set, $\mathcal{D}$, for which to obtain the corresponding labels or responses, $y_i$. That is, given the matrix $\mathbf{\Phi}_{:,I_b}$, in order to estimate the coefficient vector, $x$, we require information about a subset of $n \ll N$ components in $\{y_i\}_{i=1}^N$. The index locations of these values are contained in the set $I_l \subset \{1, \ldots, N\}$, and the resulting linear system, $y_{I_l} = \mathbf{\Phi}_{I_l,I_b}x + \eta\mathbf{1} + \epsilon$ can be solved for $x$ and $\eta$. As obtaining this information can be costly or invasive, we again seek an efficient approach for performing this task. As with basis selection discussed above, we use the differential entropy of the posterior of $x$ to construct the index set $I_l$. These values define the observations in $\mathcal{D}$ for which the corresponding label/response would be the most informative for estimating $x$.

In this section we assume $\mathbf{\Phi}_{:,I_b}$ is known. We also use the following definition for the iterative approach detailed below: Let the $i^{th}$ *row vector* of $\mathbf{\Phi}_{:,I_b}$ be represented by the *column vector* $\varphi_i$, or $\varphi_i \equiv \mathbf{\Phi}^T_{i,I_b}$. We want to iteratively select vectors, $\varphi_i$, from $\mathbf{\Phi}_{:,I_b}$, and add their index values to the set $I_l^{(k)}$, which contains $k$ indices after $k$ steps in the process. The matrix $\mathbf{\Phi}_{I_l,I_b}$, therefore, contains the rows of the observations for which we obtain corresponding labels or responses, $y_{I_l}$.

With active basis selection, the objective for each iteration was to minimize $h_k(x)$ for the linear model $y = \mathbf{\Phi}_{:,I_b^{(k)}} x + \eta\mathbf{1} + \epsilon$. Though the vector $y$ was unknown, we showed how this process was still meaningful in terms of the entropy of $x$. For active measurement selection, we again seek to minimize the entropy of $x$, this time for the actual model, $y_{I_l} = \Phi_{I_l,I_b} x + \eta\mathbf{1} + \epsilon$. For the Bayesian ridge approach, the posterior covariance of $x$ is now equal to $\left(\alpha I + \sigma^{-2}\mathbf{\Phi}^T_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b}\right)^{-1}$, where $k$ again indicates the iteration number. To minimize the entropy of $x$ following the $(k+1)^{st}$ measurement, the determinant of the inverse of this posterior covariance matrix can be maximized. As with the basis selection derivation, we expand the right matrix as follows,

$$\mathbf{\Phi}^T_{I_l^{(k+1)},I_b}\mathbf{\Phi}_{I_l^{(k+1)},I_b} = \mathbf{\Phi}^T_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b} + \varphi_i\varphi_i^T \tag{24}$$

where $i$ is the proposed $(k+1)^{st}$ index value in the set $I_l^{(k+1)}$. To calculate the determinant of this matrix, we use the following equality (though the actual calculation will include $\alpha I$ and $\sigma^2$),

$$\left|\mathbf{\Phi}^T_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b} + \varphi_i\varphi_i^T\right| = \left|\mathbf{\Phi}^T_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b}\right|\left(1 + \varphi_i^T\left(\mathbf{\Phi}^T_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b}\right)^{-1}\varphi_i\right) \tag{25}$$

In general, to maximize this term we must obtain the label/response for the datum $\gamma_i$ whose corresponding $\varphi_i$ maximizes the rightmost term. Performing this calculation with $\alpha I$ and $\sigma^2$ included, this leads to the following definition.

***Optimal Measurement Selection***: Given a set of $|I_l^{(k)}| = k$ measured labels or responses, and the matrix $\mathbf{\Phi}_{I_l^{(k)},I_b}$, the observation, $\gamma_{I_l(k+1)} \in \mathcal{D}$, for which to obtain the label/response at step $k+1$ is that having the corresponding vector $\varphi_{I_l(k+1)} \equiv \mathbf{\Phi}^T_{I_l(k+1),I_b}$, where

$$I_l(k+1) = \underset{i\in\{1,\ldots,N\}\backslash I_l^{(k)}}{\arg\max}\ \varphi_i^T\left(\alpha I + \sigma^{-2}\mathbf{\Phi}^T_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b}\right)^{-1}\varphi_i \tag{26}$$

which is the index, $i \notin I_l^{(k)}$, of the vector that minimizes the posterior differential entropy of $x$ for the Bayesian ridge approach.

We discuss the learning of parameters $\alpha$ and $\sigma^2$ in a later section. We note that this result is a special case of a general measurement acquisition method discussed in [18]. Also, see [23] for a related discussion

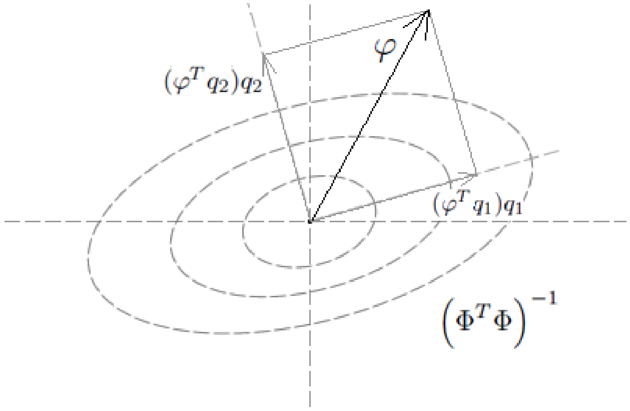


Fig. 2. The optimal measurement location at step $k+1$ is for the observation, $\gamma_i$, having the corresponding kernel vector, $\varphi_i$, that lives in the region of greatest variance, which can be measured using the eigenvectors, $q_j$, and eigenvalues or variances, $\lambda_j$, of the posterior covariance matrix of $x$ after step $k$.

in the context of compressed sensing. We again see that, though we possess the labels/responses, $y_{I_l^{(k)}}$, they do not factor into the learning process. The resulting linear model after $|I_l| = n$ steps is,

$$y_{I_l} = \mathbf{\Phi}_{I_l,I_b} x + \eta \mathbf{1} + \epsilon \tag{27}$$

where $\mathbf{\Phi}_{I_l,I_b} \in \mathbb{R}^{n\times m}$. This can be solved to obtain the best $n$-step approximation for $x \in \mathbb{R}^m$ and $\eta$.

As with active basis selection, this process has a vector space interpretation, which we show in Figure 2. From the eigendecomposition of a covariance matrix [14], generically written as $(\Phi^T\Phi)^{-1} = Q\Lambda Q^T$, the index, $I_l(k+1) = i$, of the optimal vector, $\varphi_i$, is that which maximizes

$$I_l(k+1) = \underset{i\in\{1,\ldots,N\}\backslash I_l^{(k)}}{\arg\max} \sum_{j=1}^{m} \lambda_j \left(\varphi_i^T q_j\right)^2 \tag{28}$$

where $\lambda_j$ are the eigenvalues, or variances in the directions of $q_j$, the corresponding eigenvectors of the posterior covariance matrix. The selected $\varphi_i$ is seen to be the observation whose kernel values “shoot out” into the region of greatest variance, or measure of uncertainty, of the posterior distribution of $x$.

As a termination criterion, the change in differential entropy of $x$ can again be used, this time for the true model. For a constant $\alpha$ and $\sigma^2$, this difference equals,

$$h_{k+1}(x) - h_k(x) = \frac{1}{2}\ln\left(1 + \sigma^{-2}\varphi_{I_l(k+1)}^T \left(\alpha I + \sigma^{-2}\mathbf{\Phi}_{I_l^{(k)},I_b}^T \mathbf{\Phi}_{I_l^{(k)},I_b}\right)^{-1} \varphi_{I_l(k+1)}\right) \tag{29}$$

The right term in the natural logarithm, which is the maximum of (26), is positive and monotonically decreasing to zero, as shown in the appendix. Therefore, this difference is decreasing to zero as well, and can be used to determine when enough measurements have been made by setting a threshold.

### *D. A Gaussian Process View of Active Measurement Selection*

We briefly discuss the relationship of this active learning approach to the Gaussian process [21], which will provide an alternate view of the active measurement selection process. Given the basis vectors, $\mathbf{\Phi}_{:,I_b}$, the Gaussian process collapses the hierarchical structure,

$$y \sim \mathcal{N}(\mathbf{\Phi}_{:,I_b} x, \sigma^2 I) \tag{30}$$

$$x \sim \mathcal{N}(0, \alpha^{-1} I) \tag{31}$$

by integrating out $x$ to obtain the probability $p(y|\mathbf{\Phi}_{:,I_b}, \sigma^2, \alpha)$. This distribution is Gaussian as well, with zero mean and covariance matrix $\Sigma_{\text{GP}} = \sigma^2 I + \alpha^{-1}\mathbf{\Phi}_{:,I_b}\mathbf{\Phi}_{:,I_b}^T$. By separating the data into measured and unmeasured sections, where we define the set $I_u \equiv \{1, \ldots, N\}\backslash I_l$ to contain the indices of the unmeasured locations, this distribution is,

$$\left[\begin{array}{c} y_{I_l} \\ y_{I_u} \end{array}\right] \sim \mathcal{N}\left(0, \left[\begin{array}{cc} \sigma^2 I + \alpha^{-1}\mathbf{\Phi}_{I_l,I_b}\mathbf{\Phi}_{I_l,I_b}^T & \alpha^{-1}\mathbf{\Phi}_{I_l,I_b}\mathbf{\Phi}_{I_u,I_b}^T \\ \alpha^{-1}\mathbf{\Phi}_{I_u,I_b}\mathbf{\Phi}_{I_l,I_b}^T & \sigma^2 I + \alpha^{-1}\mathbf{\Phi}_{I_u,I_b}\mathbf{\Phi}_{I_u,I_b}^T \end{array}\right]\right) \tag{32}$$

We can manipulate this distribution to give $p(y_{I_u}|y_{I_l}, \mathbf{\Phi}_{:,I_b}, \sigma^2, \alpha)$, which is again Gaussian [2]. The mean of this Gaussian serves as the prediction of $y_{I_u}$, while the covariance expresses the uncertainty in this prediction. This covariance matrix is equal to,

$$\Sigma_{y_{I_u}} = \sigma^2 I + \alpha^{-1}\mathbf{\Phi}_{I_u,I_b}\mathbf{\Phi}_{I_u,I_b}^T - \alpha^{-1}\mathbf{\Phi}_{I_u,I_b}\mathbf{\Phi}_{I_l,I_b}^T(\sigma^2 I + \alpha^{-1}\mathbf{\Phi}_{I_l,I_b}\mathbf{\Phi}_{I_l,I_b}^T)^{-1}\mathbf{\Phi}_{I_l,I_b}\mathbf{\Phi}_{I_u,I_b}^T\alpha^{-1} \tag{33}$$

A possible active learning algorithm using the Gaussian process is to measure the $y_{I_u(i)}$ whose value we are the least certain in, or the point with the largest variance along the diagonal of $\Sigma_{y_{I_u}}$. Defining $\varphi_i \equiv \mathbf{\Phi}_{i,I_b}^T$ as in the previous section, and considering only the diagonal elements of (33), then following the $k^{th}$ measurement, the index of the $(k+1)^{st}$ measurement is,

$$I_l(k+1) = \underset{i \in I_u}{\arg\max}\ \varphi_i^T\varphi_i - \varphi_i^T\mathbf{\Phi}_{I_l^{(k)},I_b}^T\left(\alpha\sigma^2 I + \mathbf{\Phi}_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b}^T\right)^{-1}\mathbf{\Phi}_{I_l^{(k)},I_b}\varphi_i \tag{34}$$

This can be connected with Section III-C by first noticing that multiplying equation (26) by $\alpha$ does not impact the selected index. In doing so, if we then expand (26) using the matrix inversion lemma (given in the appendix for reference), we see that the functions to be optimized in (26) and (34) are the same. We also note the similarity between (34) and the basis selection function in (21), which become equivalent as $\alpha\sigma^2 \rightarrow 0$. In this case, the same interpretation for basis selection holds for measurement selection when $|I_l| < |I_b|$, this time for the row space of $\mathbf{\Phi}_{:,I_b}$. In fact, as equation (34) shows, if we let $\alpha\sigma^2 \rightarrow 0$ for the first $k = |I_b|$ measurements, we can simply perform basis selection for the matrix $\mathbf{\Phi}_{:,I_b}^T$. We also observe that equations (34) and (26) suggest that $\alpha$ and $\sigma^2$ can be avoided entirely by letting $\alpha\sigma^2 \to 0$ and performing (34) for the first $k = |I_b|$ measurement locations and (26) for the remainder. However, through experimental results, we have found that additional inference for $\alpha$ and $\sigma^2$ can improve the predictive performance of the learned $x$ when $|I_l| > |I_b|$.

### *E. Learning Additional Model Parameters*

In this section, we discuss a method for learning the parameters $\alpha$, $\sigma^2$ and $\eta$. In keeping with our Bayesian analysis of the problem, we introduce additional prior distributions on $\alpha$, $\sigma^{-2}$ and $\eta$ to adaptively infer these values.

For the precision parameter, $\alpha$, we can use a conjugate gamma prior, $\mathrm{Ga}(\alpha|a/2, b/2)$, which has the analytically calculable posterior, $\mathrm{Ga}(\alpha|a', b')$, where $a' = \frac{1}{2}(a + |I_b|)$ and $b' = \frac{1}{2}(b + x^T x)$. We note that the posterior expectation of $\alpha$ is,

$$\mathbb{E}[\alpha|x] = \frac{a + |I_b|}{b + x^T x} \tag{35}$$

where $|I_b| = m$ is the dimensionality of $x$ resulting from the basis selection process. If we let $a \to 0$ and $b = |I_b|$, this will result in the posterior mean $\mathbb{E}[\alpha|x] = \frac{|I_b|}{|I_b| + x^T x}$, which restricts $\mathbb{E}[\alpha|x] < 1$. We recall that ridge regression introduced $\alpha$ as a remedy for cases where the matrix $\mathbf{\Phi}^T\mathbf{\Phi}$ has eigenvalues that are extremely small. The basis selection process of Section III-B was explicitly designed to avoid such a situation.

The inverse of the noise variance, $\sigma^{-2}$, can be given a gamma prior as well, $\mathrm{Ga}(\sigma^{-2}|c/2, d/2)$, with the analytically calculable posterior gamma distribution having parameters $c' = \frac{1}{2}(c + |I_l|)$ and $d' = \frac{1}{2}(d + \|y_{I_l} - \eta\mathbf{1} - \mathbf{\Phi}_{I_l,I_b}x\|_2^2)$. The posterior expectation is then,

$$\mathbb{E}[\sigma^{-2}|y_{I_l}, \mathbf{\Phi}_{I_l,I_b}, x] = \frac{c + |I_l|}{d + \|y_{I_l} - \eta\mathbf{1} - \mathbf{\Phi}_{I_l,I_b}x\|_2^2} \tag{36}$$

where $|I_l|$ is the dimensionality of $y_{I_l}$, or the number of measurements at the current iteration. For this prior distribution, the parameters can be set to small values. For example, setting $c = 2$ and $d \to 0$ will remove terms from calculations given below.

For the model bias, $\eta$, we use a conjugate normal prior, $\eta \sim \mathcal{N}(0, \alpha_\eta^{-1})$, and allow $\alpha_\eta \to 0$ to make this prior noninformative. The posterior distribution is Gaussian with mean, $m' = (\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T\left(y_{I_l} - \mathbf{\Phi}_{I_l^{(k)},I_b}x\right)$, and variance, $\alpha_\eta'^{-1} = \sigma^2(\mathbf{1}^T\mathbf{1})^{-1}$. We note that $\mathbf{1}^T\mathbf{1} = |I_l|$.

From a regularization perspective, the function to be optimized including all prior distributions is,

$$\begin{aligned}
\ln p(x,\alpha,\sigma^{-2},\eta|y_{I_l},\mathbf{\Phi}_{I_l,I_b},\alpha_\eta,a,b,c,d) &\propto \ln p(y_{I_l}|x,\mathbf{\Phi}_{I_l,I_b},\sigma^{-2})p(x|\alpha)p(\alpha|a,b)p(\sigma^{-2}|c,d)p(\eta|\alpha_\eta) \\
&\propto |I_l|\ln\sigma^{-2}-\sigma^{-2}\|y_{I_l}-\eta\mathbf{1}-\mathbf{\Phi}_{I_l,I_b}x\|_2^2+|I_b|\ln\alpha-\alpha x^Tx \\
&\quad +(a-2)\ln\alpha-b\alpha+(c-2)\ln\sigma^{-2}-d\sigma^{-2}-\alpha_\eta\eta^2 \qquad (37)
\end{aligned}$$

In the overdetermined case (the underdetermined case is discussed shortly) the following iterations can be made to converge to a locally optimal solution after the $k^{th}$ measurement. We note that, to preserve the posterior covariance matrix, these parameters are not integrated out as discussed in [10].

$$x = \left(\alpha\sigma^2 I+\mathbf{\Phi}^T_{I_l^{(k)},I_b}\mathbf{\Phi}_{I_l^{(k)},I_b}\right)^{-1}\mathbf{\Phi}^T_{I_l^{(k)},I_b}\left(y_{I_l^{(k)}}-\eta\mathbf{1}\right) \qquad (38)$$

$$\alpha = \frac{|I_b|}{|I_b|+x^Tx} \qquad (39)$$

$$\sigma^{-2} = \frac{|I_l^{(k)}|}{\|y_{I_l^{(k)}}-\eta\mathbf{1}-\mathbf{\Phi}_{I_l^{(k)},I_b}x\|_2^2} \qquad (40)$$

$$\eta = \frac{\mathbf{1}^T\left(y_{I_l^{(k)}}-\mathbf{\Phi}_{I_l^{(k)},I_b}x\right)}{|I_l^{(k)}|} \qquad (41)$$

where we have set $a, c = 2$ and $d, \alpha_\eta \to 0$ to remove additional terms, and $b = |I_b|$ as mentioned above. When these parameters have converged to a local optimal solution, the values for $\alpha$ and $\sigma^2$ can be used in function (26) to obtain the next measurement location, and all parameter values can be used as initial values for the next iteration. Using these prior distributions, we observe that measurement selection at step $k+1$ is no longer independent of $y_{I_l^{(k)}}$.

### *F. An Active Learning Algorithm for Kernel-Based Linear Regression and Classification*

In the outline below (Algorithm 1), we collect the above steps into an algorithmic procedure for efficiently finding the matrix $\mathbf{\Phi}_{I_l,I_b}$ and learning the resulting coefficient vector $x \in \mathbb{R}^{|I_b|}$ and parameters $\alpha$, $\sigma^2$ and $\eta$. The inputs required are initial values for $\alpha$ and $\sigma^2$ (e.g., $10^{-6}$), two thresholds and the complete matrix of kernel values, $\mathbf{\Phi}$.

Of particular interest in Algorithm 1 is the handling of $x$, $\alpha$, $\sigma^2$ and $\eta$ before the linear system becomes overdetermined. In this case, the iterations of equations (38)-(41) are unnecessary, since the maximum of the objective function (37) can be made to diverge to infinity. This is done by first finding the minimum $\ell_2$-norm solution of $x$ and then letting $\sigma^2 \to 0$. The minimum $\ell_2$ solution arises by noticing that, while (37) is diverging to infinity, it is nevertheless larger when $\|x\|_2^2$ is smaller. This is also seen in the iterations of (38)-(41) where if $\sigma^2$ is allowed to go to zero, the value of (38) approaches the minimum $\ell_2$ solution

**Algorithm 1** Active Learning for $y = \mathbf{\Phi} x + \epsilon$

```
Require: $\mathbf{\Phi}, \alpha, \sigma^2 \to 0$, Thresh_Basis, Thresh_Measure

$k = 0,\ t_b(0) \to \infty$
while $t_b(k) >$ Thresh_Basis do
  $I_b(k+1) = \arg\max_{i \in \{1,\dots,N\}}\ \phi_i^T \phi_i - \phi_i^T \mathbf{\Phi}_{:,I_b^{(k)}} \left( \mathbf{\Phi}_{:,I_b^{(k)}}^T \mathbf{\Phi}_{:,I_b^{(k)}} \right)^{-1} \mathbf{\Phi}_{:,I_b^{(k)}}^T \phi_i$
  $t_b(k+1) = \|\phi_{I_b(k+1)}^{\perp}\|_2^2$ or $\kappa^{-1} \left( \mathbf{\Phi}_{:,I_b^{(k+1)}}^T \mathbf{\Phi}_{:,I_b^{(k+1)}} \right)$
  $k \leftarrow k+1$
end while

$k = 0,\ I_l^{(0)} \equiv \emptyset,\ t_l(0) \to \infty$
while $t_l(k) >$ Thresh_Measure do
  $I_l(k+1) = \arg\max_{i \in \{1,\dots,N\} \setminus I_l^{(k)}}\ \varphi_i^T \left( \alpha I + \sigma^{-2} \mathbf{\Phi}_{I_l^{(k)},I_b}^T \mathbf{\Phi}_{I_l^{(k)},I_b} \right)^{-1} \varphi_i$
  $t_l(k) = \varphi_{I_l(k+1)}^T \left( \alpha I + \sigma^{-2} \mathbf{\Phi}_{I_l^{(k)},I_b}^T \mathbf{\Phi}_{I_l^{(k)},I_b} \right)^{-1} \varphi_{I_l(k)}$
  $k \leftarrow k+1$
  if $|I_l| \leq |I_b|$ then
    $\eta = \left( \mathbf{1}^T \left( \mathbf{\Phi}_{I_l^{(k)},I_b} \mathbf{\Phi}_{I_l^{(k)},I_b}^T \right)^{-1} y_{I_l^{(k)}} \right) \left( \mathbf{1}^T \left( \mathbf{\Phi}_{I_l^{(k)},I_b} \mathbf{\Phi}_{I_l^{(k)},I_b}^T \right)^{-1} \mathbf{1} \right)^{-1}$
    $x = \mathbf{\Phi}_{I_l^{(k)},I_b}^T \left( \mathbf{\Phi}_{I_l^{(k)},I_b} \mathbf{\Phi}_{I_l^{(k)},I_b}^T \right)^{-1} \left( y_{I_l^{(k)}} - \eta \mathbf{1} \right)$
  else
    while parameters have not converged do
      $x = \left( \alpha \sigma^2 I + \mathbf{\Phi}_{I_l^{(k)},I_b}^T \mathbf{\Phi}_{I_l^{(k)},I_b} \right)^{-1} \mathbf{\Phi}_{I_l^{(k)},I_b}^T \left( y_{I_l^{(k)}} - \eta \mathbf{1} \right)$
      $\alpha = \frac{|I_b|}{|I_b| + x^T x}$
      $\sigma^{-2} = \frac{|I_l^{(k)}|}{\|y_{I_l^{(k)}} - \eta \mathbf{1} - \mathbf{\Phi}_{I_l^{(k)},I_b} x\|_2^2}$
      $\eta = \frac{\mathbf{1}^T \left( y_{I_l^{(k)}} - \mathbf{\Phi}_{I_l^{(k)},I_b} x \right)}{|I_l^{(k)}|}$
    end while
  end if
end while
return $I_l,\ I_b,\ \alpha,\ \sigma^2,\ y_{I_l}$
```

in the limit. The specific value of $\eta$ to use in the minimum $\ell_2$ solution can be found by differentiating the $\ell_2$ norm of $x$ with respect to $\eta$, generically written as

$$\frac{\partial x^T x}{\partial \eta} = \frac{\partial}{\partial \eta}(y - \eta \mathbf{1})^T (\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}(y - \eta \mathbf{1}) \tag{42}$$

and setting to zero. This results in the value,

$$\eta = \frac{\mathbf{1}^T (\mathbf{\Phi}\mathbf{\Phi}^T)^{-1} y}{\mathbf{1}^T (\mathbf{\Phi}\mathbf{\Phi}^T)^{-1} \mathbf{1}} \tag{43}$$

which is written with the relevant subscripts in Algorithm 1. Because $\sigma^2 \rightarrow 0$, and therefore $\alpha\sigma^2 \rightarrow 0$ in the underdetermined case, the interpretation of active measurement selection given at the end of Section III-D now applies, and active label/response acquisition for the first $k = |I_b|$ measurements can be viewed as basis selection for the matrix $\mathbf{\Phi}^T_{:,I_b}$. In Algorithm 1, we do not differentiate between these two phases to save space, but note that very small initial values of $\alpha$ and $\sigma^2$ will make this statement effectively true as presented in Algorithm 1.

As mentioned in the introduction, a probit model [20] can be used to learn $x$ for the classification problem, rather than regressing directly on the labels. In this case, the vector $y_{I_l}$ is modeled as a latent variable generated from a multivariate Gaussian, with the sign of each value in $y_{I_l}$ indicating class membership for the corresponding observation (a nonzero threshold can also be used). This change to the model does not affect basis selection, since the basis functions were selected independently of $y$. The analysis for optimal measurement selection is also unchanged, though the order in which labels are obtained could be impacted in the overdetermined phase of model learning when inference for $\alpha$ and $\sigma^2$ is performed.

## IV. Applications to Regression and Classification

We demonstrate the proposed active learning algorithm on six data sets, three for classification and three for regression, as well as a synthetic data set for classification. We use a Gaussian kernel with adaptive widths,

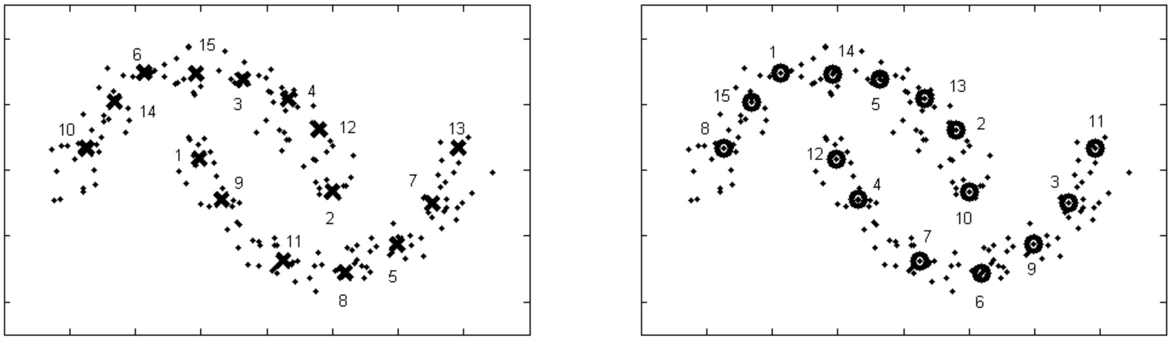
$$K(\gamma_i, \gamma_j) = \exp\left[\frac{-\|\gamma_i - \gamma_j\|_2^2}{\upsilon_i \upsilon_j}\right] \tag{44}$$

where the width value $\upsilon_i$ is associated with observation $\gamma_i$ and is set to the $5\%$ quantile of all distances to the $i^{th}$ data point. This choice is motivated by our past empirical experience with this kernel width setting, and the resulting kernel is shared by all methods compared below. Though methods exist to learn these parameters [2], we do not discuss how they can be extended to basis selection in this paper.

### *A. A Classification Example on a Synthetic Data Set*

In this section, we show the specific observations that are selected as basis functions and measurement locations for a synthesized data set and compare the selected basis functions with those chosen by other algorithms that require access to all measurements. We first generated 100 observations from each of two classes, which were used by all algorithms. This data is plotted in the figures below, where the classes are defined by the two manifolds (which we observe cannot be linearly separated in this two-dimensional space). For active learning, we show in Figure 3a the first 15 basis function locations and, given these locations, we show the first 15 measurement locations in Figure 3b. In both cases, we see that the algorithm selects points in a way that efficiently represents the two manifolds.

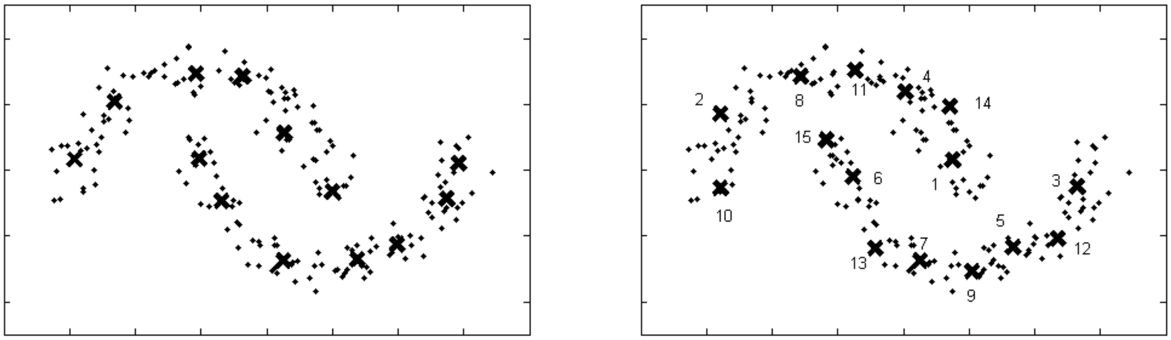
To compare, we built classifiers on the complete data set using the RVM [25] probit classifier, kernel matching pursuits [27] and the support vector machine [4]. We see in Figure 4a that the RVM converged to 13 basis functions. In Figure 4b, we show the first 15 basis functions selected by KMP, which required at least the first 8 to separate the two classes, though a predefined stopping criterion may determine that more be selected. Not shown are the results for the SVM, which selected 76 basis functions to represent the data set. Though our basis selection algorithm required no labels, we see that it performs similarly to the RVM and KMP in the way in which it selects data points to represent each manifold.



(a)

(b)

Fig. 3. (a) Basis locations for the method of Section III-B. (b) Measurement locations for the method of Section III-C using the basis functions of the left figure. Though these measurement locations correspond to the basis locations, we have empirically found that this is not always the case. Both plots show the order of selection.



(a) (b)

Fig. 4. (a) The 13 basis locations selected by the RVM. (b) The first 15 basis locations selected by KMP. The first 8 are the minimum number required to linearly separate the two classes. The SVM (not displayed) selected 76 basis functions for this data set.

### *B. Regression and Classification Examples on Real Data Sets*

We next analyze six data sets, three for regression and three for classification, and compare with other active learning approaches [1]. Below, we list all methods compared against for regression and classification. These methods consider different basis definitions, different active learning (measurement) methods, as well as a model based on random walks on a graph and one based on an alternate construction of the Gaussian process.

*Basis Comparisons* - For active learning with linear models, we consider four basis definitions for both regression and classification:

1) The original data (feature vectors) without the kernel.
2) Basis selection using a kernel as discussed in Section III-B.
3) The entire kernel, essentially treating the kernel as the original data.
4) The eigenvector basis of the kernel (or kernel PCA [2]). This last method finds the eigendecomposition of the kernel, $\mathbf{\Phi} = Q\Lambda Q^T$, and projects onto the first $m$ eigenvectors, $\mathbf{\Phi}Q_{:,1:m}$, to find an $m$-dimensional subspace in which to build a linear model. Viewing the rows of $\mathbf{\Phi}$ as vectors in $\mathbb{R}^N$, this method finds the subspace in $\mathbb{R}^m$ that preserves the greatest amount of variance in the data.

[1] All data sets used not available from the UCI database can be downloaded at www.ee.duke.edu/∼jwp4/ActLearn_IEEE_SP

*Measurement Selection Comparisons* - For active learning of regression models, we only consider the method described in Section III-C. For classification, we consider the following methods:

1) Error Reduction (ErrRed) [22]: Select the point whose value will produce the smallest average entropy of all remaining, unlabeled data. Because the label of a candidate location is unknown *a priori*, calculate the average entropy for both possible labels and take a weighted average of these two values, with the weights equal to the probability of each label for the candidate measurement location as output by the current model. For the linear models considered here, we use the logistic sigmoid function, $\sigma(y^*) = (1 + \mathrm{e}^{-y^*})^{-1}$, to convert the real-valued label prediction, $y^*$, to a value that can be interpreted as a probability.
2) Maximum Uncertainty (MaxUnc): This method selects the point whose value is the most uncertain. For linear models, this is the location that is closest to the separating hyperplane using the basis of interest, or $I_l(k+1) = \arg\min_{i \in \{1,\ldots,N\} \setminus I_l^{(k)}} \; |\varphi_i^T x^{(k)}|$, where $x^{(k)}$ is the vector of coefficients learned after measurement $k$.
3) Minimum Entropy (MinEnt): The method of Section III-C.
4) Random measurement selection: As a baseline comparison for classification, we randomly select locations for measurement. Randomly selecting measurement locations for regression was very unstable, and the results were significantly worse than for active learning. We therefore omit these results.

In addition, we consider three other active learning methods, the first for classification only and the other two for both regression and classification:

1) A Gaussian Random Field Classifier (RandWalk) [31], [32]: This approach normalizes the rows of the kernel to sum to one. The rows corresponding to the labeled data have their values replaced by zeros, except for a one in the diagonal position. This produces a random walk matrix with Dirichlet boundary conditions at the labeled locations; randomly walking to a labeled data point effectively terminates the walk because the self-transition probability at this point is equal to one. The probability of a given label for an unlabeled location is equal to the probability of terminating on a point having that particular label when the random walk is initiated from the unlabeled location of interest. The location of the next measurement is found using the error reduction criterion. Analytical calculations of all values can be derived.

2) The Gaussian Process (GP) [21]: This approach differs from Section III-D in that the matrix $\mathbf{\Phi}_{:,I_b}\mathbf{\Phi}_{:,I_b}^T$ is replaced by the kernel, $\mathbf{\Phi}$. The implied linear model prior to the marginalization of $x$ now exists in a continuous reproducing kernel Hilbert space, meaning $x$ is a function, rather than a vector in $\mathbb{R}^d$. For active learning, we use the maximum variance criterion described in Section III-D, though other active learning methods for the GP have also been proposed [15].
3) Using the full kernel for active measurement selection, we then use the RVM to learn a sparse model only on the measured data. This is called MinEnt-RVM, and is contrasted with learning the minimum $\ell_2$-norm solution for the entire kernel using the measured data, which we call MinEnt-Kernel. The measurement selection process is identical for both methods (the entire kernel is used), and these two methods for learning $x$ are meant compare the two extremes of sparseness.

*Regression Experiments* - We first present results for regression on the following three data sets:

1) The concrete data set from the UCI database. Consists of 1030 observations in $\mathbb{R}^8$ and a corresponding real-valued output that measures the compressive strength (in MPa) of a concrete mixture as a function of the observed eight ingredients.
2) The abalone data set from the UCI database. Consists of 4177 observations in $\mathbb{R}^7$ and an integer-valued output between 1 and 29 corresponding to age. Because a positive, real-valued output is interpretable for this data set, we treat this as a regression problem.
3) A data set constructed by the authors that uses the batting statistics[2] of 1037 baseball players having at least 4000 at bats. Feature vectors in $\mathbb{R}^7$ are calculated from career totals and are equal to [2B HR RBI BB R SO SB]/AB, with the response being career batting average. Note that the features were created such that the response value is not a linear function of any subset of features.

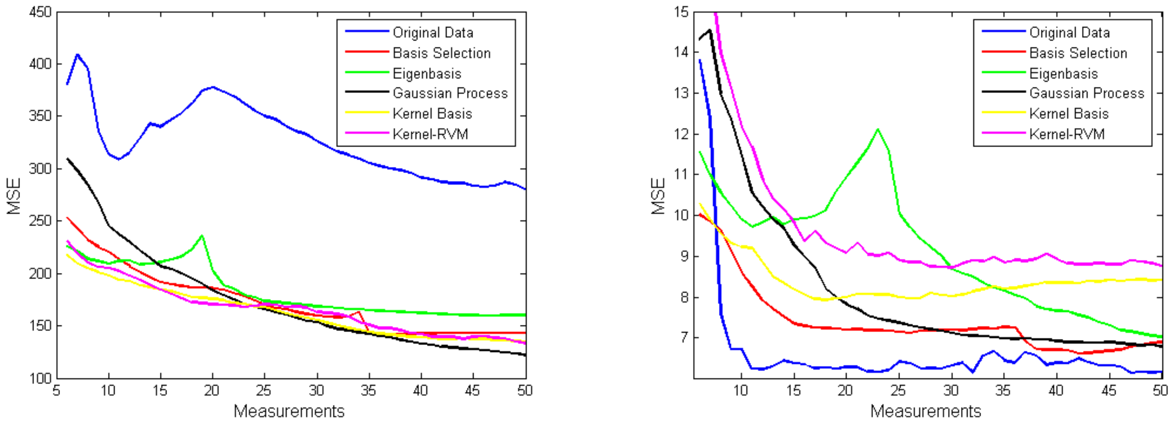
Of the above, the first two can be considered candidates for active learning, since calculating the strength of different concrete mixtures requires the destruction of materials, and determining the age of an abalone is considered a very time consuming task (according to the readme file accompanying the data). The third is a data set synthesized from real data, and is only created to further analyze performance.

In Table I, we show results for the three regression tasks using the methods described above. For each test we ran 50 trials, initializing by randomly selecting two points, followed by active learning. We use the mean square error (MSE) of all unmeasured responses as a performance measure. We average the

[2] Obtained from baseball1.com/statistics

MSE over measurement numbers 6 through 50 for each of the 50 trials and then calculate the mean and standard deviation of this average value, which is shown in Table I. In Figure 5, we plot the MSE averaged over each measurement number for the concrete and abalone data sets. For clarity, we do not plot error bars. For the regression problem, we see that the proposed method is competitive with other approaches involving a kernel, though the original data outperforms these methods for the abalone data set. We also note that the proposed method has the best performance on the baseball data set.

To select the number of basis functions for active basis selection, we used the inverse condition number stopping criterion and set a threshold of 0.01. This resulted in the following numbers of basis functions, or relevant data observations: concrete - 34, abalone - 36, baseball - 28. For the eigenbasis, we looked at the corresponding eigenvalues as well as the projected features and truncated at a point where we believed the structure of the data set was preserved, which resulted in the following basis dimensionalities: concrete - 19, abalone - 23, baseball - 18. Since these two specific thresholds may be considered arbitrary (though some value must be set), we assess the impact of basis dimensionality for these two methods in Figure 6 on the concrete and abalone data sets. These plots contain the corresponding values in Table I as a function of basis dimension. The results do not indicated an ideal stopping criterion; for the concrete data set, both methods are relatively insensitive to an increasing dimensionality, while there seems to be an optimal window for the abalone data set, which is fairly wide for basis selection.



(a)

(b)

Fig. 5. The MSE as a function of measurement number averaged over 50 runs for (a) the concrete data set and (b) the abalone data set. For clarity, the error bars are not shown, but a sense of their magnitudes is given in Table I. As is evident in plot (b), sometimes a kernel is unnecessary.

| Response-Model | Concrete | Abalone | Baseball $(\times 10^{-4})$ |
|---|---|---|---|
| MinEnt-Data | 326 ± 39 | 6.67 ± 0.53 | 4.14 ± 0.36 |
| MinEnt-BasSel | 172 ± 17 | 7.37 ± 0.62 | 3.88 ± 0.25 |
| MinEnt-EigBas | 172 ± 13 | 8.82 ± 0.87 | 4.16 ± 0.34 |
| MinEnt-Kernel | 163 ± 9 | 8.40 ± 0.56 | 3.90 ± 0.36 |
| MinEnt-RVM | 166 ± 10 | 9.78 ± 0.67 | 9.01 ± 4.60 |
| MinEnt-GP | 174 ± 11 | 8.20 ± 0.77 | 4.28 ± 0.46 |

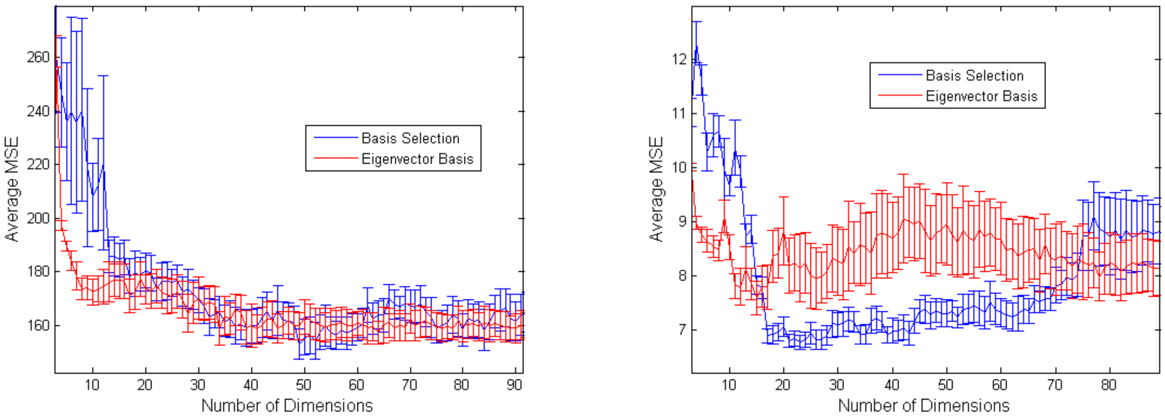
TABLE I

THE MEAN AND STANDARD DEVIATION OF THE AVERAGE MSE FROM 6TH TO THE 50TH RESPONSE MEASUREMENT. CALCULATED FROM 50 RUNS.

| Label-Model | WDBC | UXO | ION |
|---|---|---|---|
| ErrRed-Data | 0.948 ± 0.029 | 0.943 ± 0.034 | 0.676 ± 0.064 |
| MaxUnc-Data | 0.959 ± 0.021 | 0.918 ± 0.038 | 0.696 ± 0.041 |
| MinEnt-Data | 0.971 ± 0.002 | 0.912 ± 0.001 | 0.701 ± 0.009 |
| Rand-Data | 0.915 ± 0.022 | 0.887 ± 0.036 | 0.701 ± 0.034 |
| ErrRed-BasSel | 0.990 ± 0.003 | 0.972 ± 0.004 | 0.970 ± 0.007 |
| MaxUnc-BasSel | 0.989 ± 0.004 | 0.975 ± 0.002 | 0.961 ± 0.021 |
| MinEnt-BasSel | 0.986 ± 0.001 | 0.978 ± 0.000 | 0.971 ± 0.001 |
| Rand-BasSel | 0.978 ± 0.007 | 0.971 ± 0.003 | 0.945 ± 0.019 |
| ErrRed-EigBas | 0.900 ± 0.089 | 0.788 ± 0.227 | 0.880 ± 0.074 |
| MaxUnc-EigBas | 0.970 ± 0.034 | 0.931 ± 0.041 | 0.953 ± 0.018 |
| MinEnt-EigBas | 0.979 ± 0.001 | 0.977 ± 0.000 | 0.962 ± 0.002 |
| Rand-EigBas | 0.967 ± 0.024 | 0.944 ± 0.043 | 0.923 ± 0.034 |
| MinEnt-Kernel | 0.985 ± 0.004 | 0.973 ± 0.005 | 0.970 ± 0.008 |
| MinEnt-RVM | 0.971 ± 0.007 | 0.964 ± 0.006 | 0.917 ± 0.020 |
| ErrRed-RandWalk | 0.973 ± 0.014 | 0.977 ± 0.003 | 0.923 ± 0.015 |
| MinEnt-GP | 0.983 ± 0.005 | 0.974 ± 0.004 | 0.962 ± 0.012 |

TABLE II

THE MEAN AND STANDARD DEVIATION OF THE AVERAGE AUC FROM 6TH TO THE 50TH LABEL ACQUISITION. CALCULATED FROM 20 RUNS FOR ACTIVE LABEL ACQUISITION AND 100 RUNS FOR RANDOM LABEL ACQUISITION.



(a)

(b)

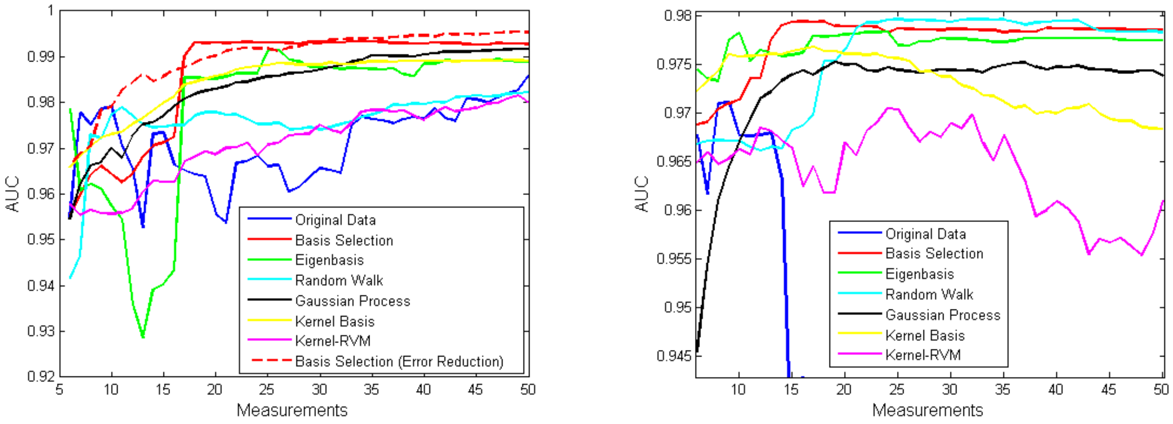
Fig. 6. The mean and standard deviation of the MSE averaged over measurement numbers 6 to 50 as a function of basis dimension for (a) the concrete data set and (b) the abalone data set.

*Classification Experiments* - We next present results for classification on the following three data sets:

1) The Wisconsin diagnostic breast cancer (WDBC) data set from the UCI database. Consists of 569 observations in $\mathbb{R}^{30}$, where each observation is classified as either malignant or benign.
2) An unexploded ordnance (UXO) data set from the authors' current research. Consists of 4612 observations in $\mathbb{R}^{45}$, where each observation is classified as either a UXO or not a UXO.
3) The ionosphere data set (ION) from the UCI database. Consists of 351 observations in $\mathbb{R}^{34}$, with the class being a "good" or "bad" returned radar signal.

In Table II, we show results for these classification tasks using the methods described above. We initialize by randomly selecting two labels, one from each class, followed by active learning. Because the results for the classification problems were more consistent than the regression problems, we ran each method only 20 times. As a baseline, we compare with randomly selecting all labels, for which we ran 100 trials. We use the area under the ROC curve (AUC), calculated from the predictions on the unlabeled data, as a performance measure. We average the AUC over labels 6 to 50 for each run and show the mean and standard deviation of this value in Table II for the different methods considered. We see that the minimum entropy label selection criterion discussed in this paper always results in the smallest standard deviation, meaning that the AUC curves are the most consistent with each trial.

This table also shows that, while the label selection process discussed in this paper does not always produce the best results, the best results do occur when the basis selection method of this paper is used and a linear model is built, though several other methods also performed well. We highlight this in Figure 7a for the WDBC data set, where we see an improvement by combining the error reduction label selection criterion with basis selection. In Figure 7b, we show results for several methods on the UXO data set, where we again see the competitive performance of the proposed algorithm. Both of these figures plot the AUC as a function of measurement number averaged over the trials and use the minimum entropy label selection method of this paper unless otherwise noted. We again do not plot error bars, which were small for the minimum entropy selection criterion.

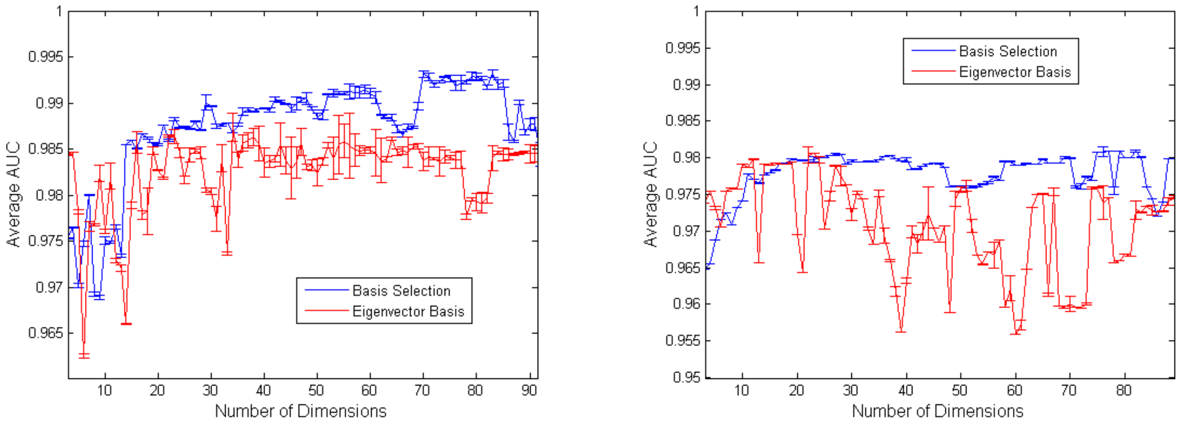


(a)

(b)

Fig. 7. The AUC as a function of measurement number averaged over 20 runs for (a) the WDBC data set and (b) the UXO data set. Plot (a) indicates that basis selection can be combined with other active learning methods to improve performance. Unless otherwise noted, all linear methods use the label selection method of this paper.

As with the regression problem, the number of dimensions to be used for the eigenbasis and basis selection must be determined in advance. We used the same criteria as discussed for regression, which for basis selection resulted in the following number of dimensions: WDBC - 22, UXO - 16, ION - 38. For the eigenbasis, these values were: WDBC - 8, UXO - 8, ION - 13. In Figure 8, we plot the mean and standard deviation of the AUC for the basis selection and eigenbasis models as a function of increasing model dimensionality (using the minimum entropy criterion). As with regression, these values correspond to those found in Table II. We observe that the eigenbasis is more volatile and does not perform as well as basis selection for the WDBC and UXO data sets. We also note that, as with the corresponding regression plots, these plots do not appear to indicate that an ideal method exists for determining how many dimensions to use in the linear model. Rather, it seems that a large window of



(a) (b)

Fig. 8. The mean and standard deviation of the AUC averaged over measurement numbers 6 to 50 as a function of basis dimension for (a) the WDBC data set and (b) the UXO data set.

dimension numbers is available for basis selection that will produce good results. The volatility of the eigenbasis dimensionality suggests that the same cannot be said for this basis.

### C. *Discussion*

We briefly discuss the relationship of four of the methods described above: basis selection, the eigenbasis (kernel PCA), the use of the entire kernel and using the kernel for measurement selection followed by the RVM; this will place basis selection in a larger context. We first observe that, as the number of dimensions increases, both basis selection and the eigenbasis converge to the entire kernel, but where the eigenbasis finds a *subspace* for the kernel features, basis selection finds a *subset* of these features. Therefore, it is reasonable that the eigenbasis required far fewer dimensions to adequately represent the data, and that performance can sometimes deteriorate when more dimensions are added, since higher dimensions contain less structure and more noise. Using the entire kernel and the minimum $\ell_2$-norm solution, we note that this approach becomes less practical as the number of observations increases, and selecting a large subset of these observations without any label or response information would call for the basis selection process discussed in this paper. We also see in the two larger data sets (the abalone data set and the UXO data set) that performance using the entire kernel deteriorates as the number of measurements increases. This is likely due to over-fitting, and suggests that basis selection is more necessary as the size of the data set increases.

The basis selection method of this paper can be viewed as a bridge between using the entire kernel, and using the entire kernel only to select measurements, followed by the RVM for subset selection. With subset selection methods such as the RVM (as well as KMP, LASSO and the SVM), the number of dimensions selected is upper bounded by the number of measurements. Therefore, when few measurements are made, the resulting model can "over-sparsify," and generalize poorly. Basis selection can prove advantageous in these early stages by preventing over-simplification of the model, while also having advantages in the later stage (larger measurement numbers) by preventing over-fitting, as the results indicate can occur when using the entire kernel. However, results for basis selection showed that the window between these two extremes of the number of basis functions appears to be large. See [23] for further discussion of the negative aspects of over-simplification in linear models.

To analyze why the kernel improves the regression and classification results for the concrete and WDBC data sets, we used the Grassberger-Procaccia (GP) algorithm [21] to estimate the intrinsic dimensionality of each data set. To review, the GP algorithm estimates the intrinsic dimensionality of a data set by measuring how the volume of a manifold within a Euclidean ball increases with respect to an increasing radius. Since the volume, $V$, of a $d$-dimensional ball is proportional to $r^d$, the intrinsic dimensionality increases linearly with a slope that equals this dimensionality, $\ln V = d \ln r + \text{const}$. We replace the volume with $C(r)$, the average fraction of points within a ball of radius $r$ centered on each point, to empirically approximate this value. We found that the intrinsic dimensionality ($iD$) of the concrete data set was $iD = 2.7$, and for the WDBC data set was $iD = 8.4$. We also performed PCA [14] to find whether this intrinsic dimensionality was linear or nonlinear, and found that the approximate minimal subspace dimensionality of the concrete data set was in $\mathbb{R}^7$, while the WDBC data set was in $\mathbb{R}^{15}$, suggesting that the respective manifolds are nonlinear.

To give a sense of how this algorithm scales, we mention that for basis selection, the WDBC data set (569 observations) required approximately 4 second to select the first 100 basis functions, while the UXO data set (4612 observations) required approximately 3 minutes on a 2.66 GHz desktop computer. The time required to add one basis function increases as the number of basis functions increases, due in large part to the increasing size of the matrix $\mathbf{\Phi}^T_{:,I_b^{(k)}}\mathbf{\Phi}_{:,I_b^{(k)}}$, the inverse of which requires on the order of $\mathcal{O}(k^3)$ operations to compute. For active measurement selection, all methods not using the error reduction criterion were comparable in speed, requiring less than 1 second to actively select 50 measurements for all data sets. The error reduction selection method (used in the linear and random walk classifiers) required significantly more time because significant computation was required to check each unlabeled data point. For the UXO data set, this time increased to approximately 15 minutes.

## V. Conclusion

In this paper, we have presented and analyzed an active learning algorithm for kernel-based linear regression and classification. For a data set, $\mathcal{D}_c = \{(\gamma_i, y_i)\}_{i=1}^N$, containing observations $\gamma_i \in \mathbb{R}^d$ and associated labels or responses, $y_i$, it is often the case that the values for $y_i$ are missing. For example, in building a model to diagnose a medical condition, we may have a large set of symptoms, $\mathcal{D} = \{\gamma_i\}_{i=1}^N$, with few or no corresponding labels $\{y_i\}_{i=1}^N$. In this situation, active learning methods can be used to select the subset of $\{y_i\}_{i=1}^N$ whose values would be the most informative for predicting the remaining, unmeasured values. However, for active learning with linear models, $y = \mathbf{\Phi}x + \eta\mathbf{1} + \epsilon$, the measure of informativeness of a given $y_i$ requires that the matrix $\mathbf{\Phi}$ be defined. When $\mathbf{\Phi}$ is constructed using a kernel function on the data, $K(\gamma_i, \gamma_j) \in \mathbb{R}$, and $\mathbf{\Phi} \in \mathbb{R}^{N \times N}$, it is generally known *a priori* that most columns of $\mathbf{\Phi}$ are irrelevant to the prediction of $\{y_i\}_{i=1}^N$. Popular methods for determining the relevant columns of $\mathbf{\Phi}$, however, are derived assuming prior knowledge of $\{y_i\}_{i=1}^N$ [4], [24], [25], [27]. Therefore, a circular dependence arises.

The active learning algorithm presented in this paper employs a greedy selection criterion to avoid this problem. Given a matrix, $\mathbf{\Phi}$, the algorithm proceeds in two steps. The first step is called basis selection, where we select columns of $\mathbf{\Phi}$, and thereby the observations in $\mathcal{D}$ on which to build a kernel. The indices of these locations are contained in the index set $I_b$. Following the selection of $m$ basis functions, this step reduces the model to $y = \mathbf{\Phi}_{:,I_b}x + \eta\mathbf{1} + \epsilon$. The second step is to sequentially obtain $n$ measurement values from the missing set $\{y_i\}_{i=1}^N$, the indices of which are contained in the index set $I_l$. This reduces the problem to the solvable model, $y_{I_l} = \mathbf{\Phi}_{I_l,I_b}x + \eta\mathbf{1} + \epsilon$. The objective function in this greedy procedure is the differential entropy of $x$. This measure arises from the Bayesian interpretation of the $\ell_2$-regularized least squares solution to $y = \mathbf{\Phi}x + \eta\mathbf{1} + \epsilon$, called ridge regression [13], and has a meaningful information-theoretic interpretation [7]. Because the posterior covariance matrix of $x$ is independent of $\{y_i\}_{i=1}^N$, basis selection can proceed without the difficulty encountered by other sparsity-promoting models. For measurement selection, inference can be performed for additional parameters in the overdetermined case, adding a dependence on measurements, $y_{I_l}$, already obtained.

In addition to active learning, we briefly mention that the basis selection process discussed in this paper has other potential uses that can motivate future research work. For example, this basis selection algorithm provides a means for performing k-means clustering along manifolds, which can possibly be used in conjunction with inversion algorithms for compressively sensed signals [8] that reside on manifolds [28], [12], [1].

# REFERENCES

[1] R.G. Baraniuk and M.B. Wakin (2009). Random projections of smooth manifolds. *Foundations of Computational Mathematics*, vol. 9, no. 1, 51-77.

[2] C. Bishop (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

[3] S. Boyd and L. Vandenberghe (2003). *Convex Optimization*, Cambridge Univ. Press, Cambridge, U.K.

[4] C.J.C. Burges (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121-167.

[5] K. Chaloner and I. Verdinelli (1995). Bayesian experimental design: A review. *Statistical Science*, 10:237-304.

[6] D.A. Cohn, Z. Ghahramani, M.I. Jordan (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129-145.

[7] T.M. Cover and J.A. Thomas (2006). *Elements of Information Theory, 2nd edition*, Wiley & Sons, New York.

[8] D.L. Donoho (2006). Compressed Sensing. *IEEE Trans. on Information Theory*, 52:1289-1306.

[9] V.V. Federov (1972). *Theory of Optimal Experiments*, Academic Press, New York.

[10] C.S. Foo, C.B. Do and A.Y. Ng (2009). A majorization-minimization algorithm for (multiple) hyperparameter learning. *The 26th International Conference on Machine Learning (ICML).*

[11] P. Grassberger and I. Procaccia (1983). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, 9:189-208.

[12] C. Hegde, M.B. Wakin and R.G. Baraniuk (2007). Random projections for manifold learning, *NIPS*.

[13] A.E. Hoerl and R.W. Kennard (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics, Special Ed., Feb. 2000*, 42(1):80-86.

[14] I. Jolliffe (2005). *Principal Component Analysis*, Wiley & Sons, New York.

[15] N.D. Lawrence, M. Seeger and R. Herbrich (2002). Fast sparse Gaussian process methods: The informative vector machine, *Advances in Neural Information Processing Systems* 15:609-616.

[16] X. Liao, Y. Zhang and L. Carin (2007). "Plan-in-advance active learning of classifiers," in A. Hero et al. (Edt), *Foundations and Applications of Sensor Management*, Springer, pages 201-220.

[17] X. Liao and L. Carin (2004). Application of the theory of optimal experiments to adaptive electromagnetic-induction sensing of buried targets. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26(8):961-972.

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

[19] D.J. MacKay (1992). The evidence framework applied to classification networks. *Neural Computation*, 4:720-736.

[20] P. McCullagh and J.A. Nelder (1989). *Generalized Linear Models, 2nd edition*, Chapman and Hall.

[21] C.E. Rasmussen and C.K.I. Williams (2006). *Gaussian Processes for Machine Learning*, MIT press.

[22] N. Roy and A. McCallum (2001). Toward optimal active learning through sampling estimation of error reduction. *The 18th International Conference on Machine Learning (ICML).*

[23] M.W. Seeger and H. Nickisch (2008). Compressed sensing and Bayesian experimental design. *The 25th International Conference on Machine Learning (ICML).*

[24] R. Tibshirani (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B*, 1-58:267-288.

[25] M.E. Tipping (2001). Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211-244.

[26] J.A. Tropp and A.C. Gilbert (2005). Signal recovery from partial information via orthogonal matching pursuit. Preprint University of Michigan.

[27] P. Vincent and Y. Bengio (2002). Kernel matching pursuit. *Machine Learning*, 48:165-187.

[28] M.B. Wakin and R.G. Baraniuk (2006). Random projections of signal manifolds. *ICASSP 2006*.

[29] Y. Zhang, X. Liao, E. Dura and L. Carin (2004). Active selection of labeled data for target detection. *Proceedings of the ICASSP*, 5:465-468.

[30] Y. Zhang, X. Liao and L. Carin (2004). Detection of buried targets via active selection of labeled data: application to sensing subsurface UXO. *IEEE Trans. on Geoscience and Remote Sensing*, 42(11):2535-2543.

[31] X. Zhu, Z. Ghahramani and J. Lafferty (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *The 20th International Conference on Machine Learning (ICML).*

[32] X. Zhu, J. Lafferty and Z. Ghahramani (2003). Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. *The 20th International Conference on Machine Learning (ICML).*

[33] H. Zou and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301 - 320.

## Appendix

1. In the appendix, we prove that the term to be maximized in (26) is monotonically decreasing in $k$ when $\alpha$ and $\sigma^2$ are constant. For clarity, we define the vectors $\psi \equiv \mathbf{\Phi}^T_{I_l(k),I_b}$ and $\varphi \equiv \mathbf{\Phi}^T_{I_l(k+1),I_b}$, being the vectors added at steps $k$ and $k+1$ respectively, and the matrix $M \equiv \left(\alpha I + \sigma^{-2}\mathbf{\Phi}^T_{I_l^{(k-1)},I_b}\mathbf{\Phi}_{I_l^{(k-1)},I_b}\right)$, being the posterior inverse covariance matrix of $x$ after measurement $k-1$. The maximum of the function in (26) at step $k+1$ is equal to $\varphi^T(M+\sigma^{-2}\psi\psi^T)^{-1}\varphi$. Using the matrix inversion lemma,

$$\begin{aligned}\varphi^T(M+\sigma^{-2}\psi\psi^T)^{-1}\varphi &= \varphi^T M^{-1}\varphi - \frac{\left(\varphi^T M^{-1}\psi\right)^2}{\sigma^2+\psi^T M^{-1}\psi} \\ &< \varphi^T M^{-1}\varphi \\ &< \psi^T M^{-1}\psi\end{aligned}$$

The first inequality arises because the rightmost term in the first line is positive. The final inequality is true by design of the iterative measurement selection process. The vector $\psi$ was selected over $\varphi$ at step $k$ because it was the vector that maximized this term.

2. For quick reference, we note that the matrix inversion lemma states the following equality:

$$\left(I + (\alpha\sigma^2)^{-1}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1} = I - \mathbf{\Phi}^T\left(\alpha\sigma^2 I + \mathbf{\Phi}\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi} \tag{45}$$