An Empirical Comparison of Spectral Learning Methods for Classification

Adam Drake and Dan Ventura
Computer Science Department
Brigham Young University, Provo, UT 84602 USA
Email: adam_drake1@yahoo.com, ventura@cs.byu.edu

***Abstract*—In this paper, we explore the problem of how to learn spectral (e.g., Fourier) models for classification problems. Specifically, we consider two sub-problems of spectral learning: (1) how to select the basis functions that will be included in the model and (2) how to assign coefficients to the selected basis functions. Interestingly, empirical results suggest that the most commonly used approach does not perform as well in practice as other approaches, while a method for assigning coefficients based on finding an optimal linear combination of low-order basis functions usually outperforms other approaches.**

I. Introduction

Spectral learning methods based on Fourier, wavelet, and other transforms have been successfully applied in both applied and theoretical domains [1], [2], [3], [4], [5], [6], [7], [8], [9], [10]. The common theme of these approaches is the end goal of representing the target function in a particular spectral representation. However, several different approaches to spectral learning have been used, and it is not clear which are most effective in typical machine learning scenarios.

In this paper, we explore the problem of how to best learn spectral representations for classification problems. In doing so, we compare and analyze new and old approaches to the two main phases of the spectral learning process: determining which basis functions to include in the model and determining the coefficients to assign to each basis function.

II. Background

Spectral representations provide an alternative representation of a function. For example, consider the Fourier spectrum. Suppose $f : \{0,1\}^n \longrightarrow \mathbb{R}$. Then the Fourier spectrum of f, denoted $\hat{f}$, is given by

$$\hat{f}(\alpha) = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} f(x)\chi_\alpha(x) \tag{1}$$

where $\alpha \in \{0,1\}^n$ labels basis function χ_α, defined as

$$\chi_\alpha(x) = \begin{cases} +1 : \text{if } \sum_i \alpha_i x_i \text{ is even} \\ -1 : \text{if } \sum_i \alpha_i x_i \text{ is odd} \end{cases} \tag{2}$$

where α_i and x_i denote the i^{th} binary digits of α and x. Each Fourier coefficient $\hat{f}(\alpha)$ corresponds to a basis function, χ_α, and the sign and magnitude of $\hat{f}(\alpha)$ indicate the correlation between f and χ_α. Large positive and negative coefficients indicate significant positive and negative correlations, respectively, while small coefficients indicate little correlation.

Any f can be recovered from its Fourier representation by:

$$f(x) = \sum_{\alpha \in \{0,1\}^n} \hat{f}(\alpha)\chi_\alpha(x) \tag{3}$$

As Eq. 3 shows, the Fourier spectrum provides a representation of f as a linear combination of the Fourier basis functions.

In the case of an n-dimensional Boolean-input function, the Fourier basis functions are XOR functions, each returning -1 iff the XOR of a particular subset of the inputs is true. The subset is implicitly defined by α in Equation 2. Since $\alpha_i x_i = 0$ when $\alpha_i = 0$ and $\alpha_i x_i = x_i$ when $\alpha_i = 1$, the output of χ_α depends only inputs for which $\alpha_i = 1$. The order of any χ_α is given by $\sum_i \alpha_i$, the number of inputs that are relevant to χ_α.

By changing the basis function definition, instead of XORs, we obtain new bases of AND (ξ) and OR (ζ) basis functions:

$$\xi_\alpha(x) = \begin{cases} +1 : \text{if } \sum_i \alpha_i x_i < \sum_i \alpha_i \\ -1 : \text{if } \sum_i \alpha_i x_i = \sum_i \alpha_i \end{cases} \tag{4}$$

$$\zeta_\alpha(x) = \begin{cases} +1 : \text{if } \sum_i \alpha_i x_i = 0 \\ -1 : \text{if } \sum_i \alpha_i x_i > 0 \end{cases} \tag{5}$$

By replacing the Fourier basis functions in Equation 1 with either of these sets of basis functions, we obtain new "correlation spectra" —the coefficients reveal the correlation between f and either the AND or OR functions, just as the Fourier coefficients do for the XOR functions. Note, however, that unlike in the XOR case, the coefficients obtained from Equation 1 will not generally give the linear combination of AND or OR functions that equals f. There is another transform equation that gives the linear combination (but not the correlation); however, only the correlation spectrum will be of interest here.

III. Spectral Learning Methods

Given a set X of $\langle x, f(x)\rangle$ examples, a spectral learning algorithm attempts to learn a spectral representation of f that approximates it well. Since the number of basis functions is exponential in the number of inputs to a function, a spectral learning algorithm will typically select a subset of basis functions to use in its model, implicitly assigning coefficients of 0 to the remaining basis functions. If A is the set of labels of basis functions included in the model, then a spectral learner's approximation of f is given by the following:

$$\tilde{f}(x) = \sum_{\alpha \in A} \hat{f}(\alpha)\phi_\alpha(x) \tag{6}$$

where ϕ_α is a general basis function reference that could be replaced by any of the basis functions defined above.

Spectral learning algorithms can be applied to Boolean classification problems by encoding the outputs of positive and negative examples as -1.0 and 1.0, respectively, and using the sign of the model's output to make classifications:

$$\tilde{f}(x) = \begin{cases} false & : \text{if } \sum_{\alpha \in A} \hat{f}(\alpha)\phi_\alpha(x) \geq 0 \\ true & : \text{if } \sum_{\alpha \in A} \hat{f}(\alpha)\phi_\alpha(x) < 0 \end{cases} \quad (7)$$

The task of a spectral learner is to determine which basis functions to include in the model and what coefficient values to assign to each basis function.

A. Selecting Basis Functions

Three approaches to basis function selection are considered in this paper: Most-Correlated, Low-Order, and AdaBoost.

1) Most-Correlated: The most common approach to basis function selection is to select the basis functions that are most correlated with f, or, equivalently, that have the largest coefficients in the correlation spectrum of f (Equation 1) [11], [1], [2], [3], [8], [10]. Although the true coefficients are unknown, they can be estimated from X:

$$\tilde{\hat{f}}(\alpha) = \frac{1}{|X|} \sum_{\langle x, f(x) \rangle \in X} f(x)\phi_\alpha(x) \quad (8)$$

Stated precisely, the Most-Correlated selection method used in this paper selects basis functions according to rule:

$$\begin{array}{l} \phi_\alpha \text{ is preferred to } \phi_\beta \text{ iff:} \\ \quad (|\hat{f}(\alpha)| > |\hat{f}(\beta)|) \vee \\ \quad\quad (|\hat{f}(\alpha)| = |\hat{f}(\beta)| \wedge \sum_i \alpha_i < \sum_i \beta_i) \end{array}$$

Note that ties in coefficient size are broken in favor of lower-order basis functions (further ties are broken randomly).

For any basis, the Most-Correlated approach makes sense from a feature selection perspective, as basis functions that are correlated with f should be better features. For a basis in which the correlation spectrum gives the representation of a function in that basis, such as the Fourier spectrum, this approach can also be motivated by the goal of trying to approximate the true representation of f in that basis, with the sensible strategy being to select the basis functions with large coefficients, as they carry the most "weight" in the linear combination.

2) Low-Order: Another reasonable approach to basis function selection is to use the low-order basis functions (e.g., to select all basis functions for which $\sum_i \alpha_i \leq k$) [12], [9]. The Low-Order approach used in this paper selects basis functions according to the following rule:

$$\begin{array}{l} \phi_\alpha \text{ is preferred to } \phi_\beta \text{ iff:} \\ \quad (\sum_i \alpha_i < \sum_i \beta_i) \vee \\ \quad\quad (\sum_i \alpha_i = \sum_i \beta_i \wedge |\hat{f}(\alpha)| > |\hat{f}(\beta)|) \end{array}$$

Note that basis functions of the same order are selected in order of highest correlation with the training data. (If there is still a tie, it is broken randomly.) Thus, the Low-Order and Most-Correlated methods both favor low-order functions and high correlations, and they differ only in which criterion is considered more important.

SelectBasisFunctions–AdaBoost(X, T)

(1) $A \leftarrow \emptyset$
(2) for each $\langle x, f(x) \rangle \in X$
(3) $\quad w_{\langle x, f(x) \rangle} \leftarrow \frac{1}{|X|}$
(4) for $t = 1$ to T
(5) $\quad X_t \leftarrow \{\langle x, w_{\langle x, f(x) \rangle} f(x) \rangle : \langle x, f(x) \rangle \in X\}$
(6) $\quad \phi_{\alpha_t} \leftarrow$ SelectCorrelatedFunction(X_t)
(7) $\quad A \leftarrow A \cup \phi_\alpha$
(8) $\quad \epsilon_t \leftarrow \sum_{\{\langle x, f(x) \rangle : \phi_{\alpha_t}(x) \neq f(x)\}} w_{\langle x, f(x) \rangle}$
(9) $\quad$ for each $\langle x, f(x) \rangle \in X$ s.t. $\phi_{\alpha_t}(x) = f(x)$
(10) $\quad\quad w_{\langle x, f(x) \rangle} \leftarrow w_{\langle x, f(x) \rangle} (\frac{\epsilon_t}{1 - \epsilon_t})$
(11) $\quad z \leftarrow \sum_{\langle x, f(x) \rangle \in X} w_{\langle x, f(x) \rangle}$
(12) $\quad$ for each $\langle x, f(x) \rangle \in X$
(13) $\quad\quad w_{\langle x, f(x) \rangle} \leftarrow w_{\langle x, f(x) \rangle} / z$
(14) return A

Fig. 1. The AdaBoost basis function selection procedure.

One motivation for preferring low-order functions is that it seems reasonable to expect that in practice f is more likely to be correlated with lower-order functions. (In fact, a low-order approach can be viewed as a most-correlated approach with the prior assumption that lower-order functions will be more correlated with f.) Low-order functions may be more correlated with f because they are defined over fewer inputs and therefore represent a simpler interaction between inputs. It also seems reasonable to expect that lower-order functions will be more likely to generalize well.

3) AdaBoost: An alternative approach to basis function selection is to select basis functions in conjunction with a boosting algorithm [5], generating an ensemble of learners that makes classifications by weighted vote. The learners are trained iteratively, typically with the first learner trained on the original data set and subsequent learners trained on weighted data sets in which examples that were misclassified by previously-trained learners receive more weight. If the learners are spectral learners whose models consist of a single basis function, then the result is just a spectral representation in which the basis functions (and possibly coefficients) were selected in conjunction with a boosting algorithm.

The AdaBoost basis function selection approach used in this paper is based on the AdaBoost.M1 algorithm [13], and is illustrated in Figure 1. In each boosting iteration t, the weights for each example are used to create a weighted data set, X_t (line 5), in which the weights are implicitly represented by converting the $f(x)$ values from ± 1 to $\pm w_{\langle x, f(x) \rangle}$. Then, a basis function that is highly correlated with X_t is selected (line 6) and added to the solution (line 7). The distribution of weight over the examples is initially uniform (line 3), but it is updated each iteration (lines 8-13) so that examples that are classified correctly by the most recently added basis function receive less weight (lines 9-10). (Note: For simplicity, the algorithm in Figure 1 is presented as if each ϕ_{α_t} is positively correlated with X_t. However, if ϕ_{α_t} is negatively correlated, each occurrence of $\phi_{\alpha_t}(x)$ in lines 8 and 9 should be replaced with $(-\phi_{\alpha_t}(x))$.)

B. Assigning Coefficients

Three methods of assigning coefficients to selected basis functions are considered in this paper: Data-Estimate, Min-Squared-Error, and AdaBoost.

1) Data-Estimate: The Data-Estimate approach, or some variation of it, is by far the most common method for assigning coefficients to basis functions [11], [12], [1], [2], [8], [9], [10]. In its basic form, each basis function is assigned the coefficient that is estimated from training data. Typically, this is done by Equation 8, which is also the method used here.

If the goal is to approximate the true spectral representation of f, then setting each coefficient to the value estimated from the training data would be a natural choice, especially if the basis function selection approach is motivated by the same goal. Regardless of how basis functions are selected, however, the Data-Estimate method can be reasonably motivated as an ensemble-building approach that weights each basis function in proportion to its classification accuracy over X.

2) Min-Squared-Error: The Min-Squared-Error coefficient assignment method can be motivated by viewing spectral learning from a feature selection perspective. Basis function selection can be thought of as the task of identifying a good set of features, while coefficient assignment can be thought of as the task of learning an "optimal" linear combination of the features, without regard to whether the resulting combination resembles the true spectral representation of the function. In the Min-Squared-Error approach, the optimal linear combination of the set A of selected basis functions is the one that minimizes the squared error over the training data:

$$\underset{\tilde{\hat{f}}(\alpha_1),\ldots,\tilde{\hat{f}}(\alpha_{|A|})}{\operatorname{argmin}} \left(\sum_{\langle x, f(x)\rangle \in X} \left(f(x) - \sum_{\alpha \in A} \tilde{\hat{f}}(\alpha)\phi_\alpha(x) \right)^2 \right)$$

Motivations for using squared error as the metric for optimality in the linear combination include the fact that it generalizes naturally to regression problems and that it is easily computed. Other common metrics, such as the number of misclassifications or the distance from the decision surface to the nearest training examples of each class are not considered here.

Note that there may not be a unique solution to the least-squares problem, indicating that with respect to the data there is redundancy in the set of selected basis functions. To resolve this issue, basis functions are considered for inclusion in the model iteratively (either in the order defined by the preference function or the order in which they were added to the model by AdaBoost), and any basis functions whose inclusion would introduce redundancy are not added to the model.

3) AdaBoost: The AdaBoost coefficient assignment method makes sense only in the context of the AdaBoost basis function selection method. In the AdaBoost.M1 algorithm, each learner is assigned a coefficient whose magnitude is proportional to the learner's accuracy on its weighted set of training data. In terms of the AdaBoost basis function selection method described in Figure 1, each coefficient is given by the following:

$$\hat{f}(\alpha_t) = \pm \log\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \tag{9}$$

where ϵ_t is the (weighted) misclassification rate of ϕ_{α_t} and the sign of $\hat{f}(\alpha_t)$ is negative iff $\sum_{\langle x, f(x)\rangle \in X_t} f(x)\phi_{\alpha_t}(x) < 0$ (i.e., if ϕ_{α_t} is negatively correlated with X_t).

IV. Empirical Results

In this section, the basis function selection and coefficient assignment methods are compared on nine Boolean classification problems [14] with each of the previously defined spectral representations (i.e., the AND, OR, and XOR bases). For each learning problem, the data was partitioned 100 times into training and test sets (with 10% used for testing), and the average classification accuracy on the test set when training on the corresponding training set was recorded. Each method used the same 100 splits of data, and results were averaged over those 100 trials. Statistically significant differences were measured pair-wise by a paired permutation test, with significant differences defined as those for which $p \leq 0.01$.

For each spectral learning approach there is a single free parameter: T, the number of basis functions to include in the model. This parameter was set automatically as part of the learning process. Specifically, each learner would split its training data into training and validation sets (with 10% held out for validation), and would then estimate its generalization performance with each number of basis functions from 1 to T_{max}. After repeating this on 10 random partitions of the training data and averaging results, T was set to the number of basis functions that maximized classification accuracy on the validation set (with ties broken in favor of fewer basis functions). Then, the learner would train on the entire training set with the selected T value.

The following sections present pair-wise comparisons of the spectral learning approaches, and tables of results focus on those cases where statistically significant differences between methods were observed. A complete listing of the results can be seen in Table VIII at the end of the paper.

A. Assigning Coefficients

Tables I and II show the average test accuracy when using the Data-Estimate and Min-Squared-Error coefficient assignment methods with Low-Order and Most-Correlated basis function selection. Of the 27 possible combinations of data set and basis, the tables show only those cases for which a statistically significant difference between methods was observed. In each case, the higher accuracy is bolded. Where there were significant differences, the Min-Squared-Error approach is usually superior to the Data-Estimate approach.

One advantage of the Min-Squared-Error approach is that it has more flexibility in modeling functions. In the Data-Estimate approach, the coefficients are assigned independently, without regard to what other basis functions may be in the model. In the Min-Squared-Error approach, the coefficients are set as a group to be "optimal" with respect to the selected set of basis functions. Of course, the minimum squared error linear combination is not certain to be better, and increased flexibility can increase the likelihood of overfitting, which may explain why the Min-Squared-Error approach occasionally performed worse. However, it seems to be a better approach in general.

Table III shows a comparison of the three coefficient assignment methods when AdaBoost is used to select basis functions. In general, the AdaBoost coefficient assignment method gives the best results.

TABLE I. COMPARISON OF COEFFICIENT ASSIGNMENT METHODS FOR THE LOW-ORDER BASIS FUNCTION SELECTION APPROACH. WHERE SIGNIFICANT DIFFERENCES WERE OBSERVED (SHOWN BELOW), THE MIN-SQUARED-ERROR APPROACH IS USUALLY SUPERIOR.

DATA SET	BASIS	DATAEST	MINSQERR
CHESS	AND	87.2%	**95.2%**
CHESS	OR	86.8%	**96.6%**
CHESS	XOR	87.0%	**94.9%**
GERMAN	AND	69.9%	**73.1%**
GERMAN	OR	70.8%	**73.7%**
GERMAN	XOR	70.4%	**73.2%**
PIMA	XOR	72.6%	**73.5%**
SPECT	AND	79.3%	**81.9%**
SPECT	OR	79.0%	**81.9%**
SPECT	XOR	77.3%	**81.6%**
VOTING	AND	95.4%	**95.8%**
VOTING	XOR	95.3%	**95.9%**
WISC1	AND	95.5%	**96.0%**
WISC1	OR	**96.3%**	95.8%
WISC1	XOR	95.6%	**96.2%**
WISC2	AND	**75.5%**	71.5%
WISC2	OR	**75.6%**	72.4%
WISC2	XOR	**76.3%**	72.6%
WISC3	AND	91.3%	**94.3%**

TABLE II. COMPARISON OF COEFFICIENT ASSIGNMENT METHODS FOR THE MOST-CORRELATED BASIS FUNCTION SELECTION APPROACH. WHERE SIGNIFICANT DIFFERENCES WERE OBSERVED (SHOWN BELOW), THE MIN-SQUARED-ERROR APPROACH IS USUALLY SUPERIOR.

DATA SET	BASIS	DATAEST	MINSQERR
CHESS	AND	80.7%	**81.0%**
CHESS	OR	77.6%	**89.0%**
CHESS	XOR	75.5%	**83.2%**
GERMAN	AND	69.6%	**70.9%**
HEART	AND	79.2%	**81.5%**
PIMA	AND	**74.1%**	73.3%
PIMA	OR	72.5%	**73.6%**
SPECT	OR	**83.7%**	82.6%
VOTING	AND	95.5%	**95.8%**
WISC1	AND	93.2%	**96.0%**
WISC3	AND	90.8%	**92.8%**

Since both the Data-Estimate and AdaBoost methods assign coefficients that are proportional to classification accuracy, with the primary difference being whether accuracy is measured with respect to the original data or a weighted version of the data, it may be surprising that AdaBoost gave a significantly better result so often. However, an important difference is that while the Data-Estimate coefficients are assigned independently, the iteratively-assigned AdaBoost coefficients are each dependent on previously-added basis functions. In AdaBoost, the weighted data sets implicitly carry information about previously added basis functions, which allows the assigned coefficients to be "optimized" in a sense with respect to previously added basis functions. As with the Min-Squared-Error approach, however, this extra flexibility may lead to overfitting in some cases.

There were few significant differences between the Min-Squared-Error and AdaBoost coefficient assignment methods. Interestingly, however, in those cases where there was a difference, the AdaBoost method is always superior.

TABLE III. COMPARISON OF COEFFICIENT ASSIGNMENT METHODS FOR THE ADABOOST BASIS FUNCTION SELECTION APPROACH. WHERE SIGNIFICANT DIFFERENCES WERE OBSERVED (SHOWN BELOW), THE ADABOOST COEFFICIENT ASSIGNMENT METHOD IS USUALLY SUPERIOR TO THE OTHER APPROACHES.

DATA SET	BASIS	ADABOOST	DATAEST
CHESS	AND	**97.6%**	94.0%
CHESS	OR	**96.1%**	93.9%
CHESS	XOR	**97.7%**	94.8%
GERMAN	AND	**72.7%**	71.7%
GERMAN	OR	**72.6%**	71.4%
HEART	AND	**81.3%**	79.1%
HEART	OR	**81.5%**	78.2%
SPECT	AND	**83.5%**	78.7%
SPECT	OR	82.3%	**83.9%**
WISC1	AND	**95.9%**	92.8%
WISC3	AND	**93.9%**	92.7%
WISC3	OR	**94.8%**	92.8%
WISC3	XOR	91.6%	**92.8%**

DATA SET	BASIS	ADABOOST	MINSQERR
CHESS	AND	**97.6%**	95.8%
CHESS	OR	**96.1%**	95.4%
CHESS	XOR	**97.7%**	96.4%
HEART	AND	**81.3%**	79.4%
WISC1	OR	**96.0%**	95.5%

TABLE IV. COMPARISON OF THE LOW-ORDER AND MOST-CORRELATED BASIS FUNCTION SELECTION METHODS. WHERE SIGNIFICANT DIFFERENCES WERE OBSERVED (SHOWN BELOW), THE LOW-ORDER APPROACH IS CONSISTENTLY SUPERIOR.

DATA SET	BASIS	LOW-ORDER	MOST-CORR
CHESS	AND	**95.2%**	81.0%
CHESS	OR	**96.6%**	89.0%
CHESS	XOR	**94.9%**	83.2%
GERMAN	AND	**73.1%**	70.9%
GERMAN	OR	**73.7%**	70.1%
GERMAN	XOR	**73.2%**	71.6%
HEART	OR	**83.2%**	78.4%
SPECT	AND	**81.9%**	77.9%
SPECT	XOR	**81.6%**	78.0%
VOTING	XOR	**95.9%**	95.4%
WISC1	OR	**95.8%**	94.9%
WISC2	AND	71.5%	**73.8%**
WISC3	AND	**94.3%**	92.8%
WISC3	OR	**94.4%**	92.8%
WISC3	XOR	**94.0%**	91.8%

B. Selecting Basis Functions

Tables IV, V, and VI provide pair-wise comparisons of the AdaBoost, Low-Order, and Most-Correlated basis function selection methods when each is combined with its preferred coefficient assignment method (i.e., Min-Squared-Error coefficient assignment for the Low-Order and Most-Correlated selection methods, and AdaBoost coefficient assignment for the AdaBoost selection method). Again, only cases for which there was a statistically significant difference between methods are shown, and the higher accuracy in each case is bolded.

Table IV reveals a clear superiority of the Low-Order approach over the Most-Correlated approach. This result is interesting, as the Most-Correlated approach would seem to have an advantage: it can select correlated basis functions

TABLE V. COMPARISON OF THE ADABOOST AND MOST-CORRELATED BASIS FUNCTION SELECTION METHODS. WHERE SIGNIFICANT DIFFERENCES WERE OBSERVED (SHOWN BELOW), THE ADABOOST APPROACH IS USUALLY SUPERIOR.

DATA SET	BASIS	ADABOOST	MOST-CORR
CHESS	AND	**97.6%**	81.0%
CHESS	OR	**96.1%**	89.0%
CHESS	XOR	**97.7%**	83.2%
GERMAN	AND	**72.7%**	70.9%
GERMAN	OR	**72.6%**	70.1%
HEART	OR	**81.5%**	78.4%
HEART	XOR	79.3%	**82.7%**
PIMA	AND	**74.0%**	73.3%
PIMA	OR	72.3%	**73.6%**
SPECT	AND	**83.5%**	77.9%
WISC1	OR	**96.0%**	94.9%
WISC1	XOR	94.8%	**95.9%**
WISC3	AND	**93.9%**	92.8%
WISC3	OR	**94.8%**	92.8%

TABLE VI. COMPARISON OF THE ADABOOST AND LOW-ORDER BASIS FUNCTION SELECTION METHODS. WHERE SIGNIFICANT DIFFERENCES WERE OBSERVED (SHOWN BELOW), THE LOW-ORDER APPROACH IS USUALLY SUPERIOR.

DATA SET	BASIS	ADABOOST	LOW-ORDER
CHESS	AND	**97.6%**	95.2%
CHESS	OR	96.1%	**96.6%**
CHESS	XOR	**97.7%**	94.9%
GERMAN	OR	72.6%	**73.7%**
GERMAN	XOR	71.8%	**73.2%**
HEART	OR	81.5%	**83.2%**
HEART	XOR	79.3%	**83.3%**
PIMA	OR	72.3%	**73.4%**
SPECT	AND	**83.5%**	81.9%
SPECT	XOR	78.7%	**81.6%**
VOTING	XOR	95.3%	**95.9%**
WISC1	XOR	94.8%	**96.2%**
WISC3	XOR	91.6%	**94.0%**

from any part of the spectrum, while the Low-Order method is restricted to low-order functions. As mentioned previously, it seems reasonable to expect that low-order functions will typically exhibit higher correlation. Thus, we might expect that the two approaches would perform similarly, as both would tend to select low-order functions. However, the results indicate that the Most-Correlated approach often selects high-order functions that are *individually* most correlated with X, but the end result is a set of functions that is *collectively* less correlated with f. Two possible reasons for this are (1) the higher-order functions do not generalize as well or (2) the most correlated basis functions are less effective as a set of features than other sets of basis functions.

Further analysis points to the second reason. Empirical results show that although low-order basis functions tend to be more correlated with X (and f) on average, the error in training-data estimates of correlation do not seem to be worse for high-order basis functions than low-order basis functions. Thus, the highly correlated high-order basis functions may not necessarily be bad features. However, the results in Table VII suggest one problem with the Most-Correlated approach. For each data set and basis combination shown in Table IV, Table VII shows the average correlation, measured over the training data, between the first 10 basis functions selected by each method. The results indicate that the most correlated basis functions tend to also be very correlated with each other. Consequently, it is likely that although the Most-Correlated approach picks basis functions that are highly correlated with X (and probably f), they may tend to be so correlated with each other that there is little gained by combining them.

TABLE VII. AVERAGE CORRELATION (OVER THE TRAINING DATA) BETWEEN THE FIRST 10 BASIS FUNCTIONS SELECTED BY EACH SELECTION METHOD. THE MOST-CORRELATED APPROACH TENDS TO SELECT BASIS FUNCTIONS THAT ARE HIGHLY CORRELATED WITH EACH OTHER, WHICH CAN HAMPER LEARNING.

DATA SET	BASIS	ADABST	LOWORD	MOSTCO
CHESS	AND	0.232	0.242	0.996
CHESS	OR	0.356	0.242	0.484
CHESS	XOR	0.156	0.240	0.992
GERMAN	AND	0.332	0.358	0.794
GERMAN	OR	0.292	0.364	0.892
GERMAN	XOR	0.106	0.322	0.818
HEART	OR	0.256	0.270	0.720
SPECT	AND	0.382	0.292	0.566
SPECT	XOR	0.104	0.286	0.290
VOTING	XOR	0.184	0.498	0.622
WISC1	OR	0.644	0.614	0.914
WISC2	AND	0.408	0.404	0.942
WISC3	AND	0.488	0.710	0.862
WISC3	OR	0.472	0.704	0.994
WISC3	XOR	0.268	0.688	0.916

The performance advantage of the AdaBoost approach over the Most-Correlated approach (Table V) is also interesting, and may also be explained in part by Table VII. Both methods are based on choosing correlated basis functions. However, AdaBoost's selection of basis functions that are correlated with weighted data sets naturally de-correlates the selected functions and results in a set of basis functions that are correlated with different local regions of f.

Finally, Table VI shows that the Low-Order approach usually outperforms the AdaBoost approach when there is a significant difference. Thus, the Low-Order approach seems to be the best approach overall. However, the AdaBoost approach gave the best result on two of the nine data sets (Chess and SPECT), and may therefore be worth considering. The Most-Correlated approach, on the other hand, would seem to be less useful to consider, as it is often worse than the others and it was never significantly better than both of the others.

V. CONCLUSION

Spectral approaches to machine learning have been successfully applied in several domains, and different methods for learning spectral representations have been proposed. In this paper, we have compared the fundamental approaches to selecting basis functions and assigning coefficients. Interestingly, empirical results suggest that the spectral learning approach that is most common, selecting the most correlated basis functions and estimating their coefficients from data, which is motivated by a desire to estimate the function's true spectral representation, may be the worst approach for typical machine learning problems. On the other hand, attempting

TABLE VIII. COMPLETE TABLE OF RESULTS FOR THE VARIOUS COMBINATIONS OF BASIS FUNCTION SELECTION METHODS (MOST-CORRELATED, LOW-ORDER, AND ADABOOST) AND COEFFICIENT ASSIGNMENT METHODS (DATA-ESTIMATE, MIN-SQUARED-ERROR, AND ADABOOST).

DATA SET	BASIS	MOSTCO+ DATAEST	MOSTCO+ MINSQERR	LOWORD+ DATAEST	LOWORD+ MINSQERR	ADABST+ DATAEST	ADABST+ MINSQERR	ADABST+ ADABST
CHESS	AND	80.7%	81.0%	87.2%	95.2%	94.0%	95.8%	97.6%
CHESS	OR	77.6%	89.0%	86.8%	96.6%	93.9%	95.4%	96.1%
CHESS	XOR	75.5%	83.2%	87.0%	94.9%	94.8%	96.4%	97.7%
GERMAN	AND	69.6%	70.9%	69.9%	73.1%	71.7%	72.2%	72.7%
GERMAN	OR	70.7%	70.1%	70.8%	73.7%	71.4%	71.9%	72.6%
GERMAN	XOR	71.4%	71.6%	70.4%	73.2%	72.1%	72.1%	71.8%
HEART	AND	79.2%	81.5%	83.2%	82.4%	79.1%	79.4%	81.3%
HEART	OR	79.3%	78.4%	83.2%	83.2%	78.2%	81.3%	81.5%
HEART	XOR	81.6%	82.7%	82.9%	83.3%	80.8%	79.4%	79.3%
PIMA	AND	74.1%	73.3%	74.1%	73.6%	74.1%	74.1%	74.0%
PIMA	OR	72.5%	73.6%	73.6%	73.4%	71.7%	72.6%	72.3%
PIMA	XOR	72.9%	73.1%	72.6%	73.5%	73.7%	72.9%	73.2%
SPECT	AND	78.1%	77.9%	79.3%	81.9%	78.7%	83.1%	83.5%
SPECT	OR	83.7%	82.6%	79.0%	81.9%	83.9%	82.1%	82.3%
SPECT	XOR	78.4%	78.0%	77.3%	81.6%	78.9%	79.1%	78.7%
VOTING	AND	95.5%	95.8%	95.4%	95.8%	95.3%	95.8%	95.8%
VOTING	OR	95.4%	95.5%	95.5%	95.7%	95.4%	95.4%	95.2%
VOTING	XOR	95.1%	95.4%	95.3%	95.9%	95.4%	95.4%	95.3%
WISC1	AND	93.2%	96.0%	95.5%	96.0%	92.8%	96.0%	95.9%
WISC1	OR	95.1%	94.9%	96.3%	95.8%	95.9%	95.5%	96.0%
WISC1	XOR	95.8%	95.9%	95.6%	96.2%	94.4%	94.6%	94.8%
WISC2	AND	74.6%	73.8%	75.5%	71.5%	73.0%	72.4%	73.1%
WISC2	OR	72.4%	72.0%	75.6%	72.4%	72.6%	71.6%	72.3%
WISC2	XOR	71.7%	71.0%	76.3%	72.6%	70.7%	70.9%	71.5%
WISC3	AND	90.8%	92.8%	91.3%	94.3%	92.7%	93.8%	93.9%
WISC3	OR	93.0%	92.8%	93.8%	94.4%	92.8%	94.3%	94.8%
WISC3	XOR	91.7%	91.8%	93.7%	94.0%	92.8%	92.3%	91.6%

to learn an optimal linear combination of low-order basis functions appears to be a more effective approach.

Although the results presented in this paper suggest that a spectral learner will perform better if it limits itself to low-order basis functions (even if there are higher-order basis functions that appear to be more highly correlated), not all functions can be approximated well by only low-order basis functions, and it seems reasonable to expect that in some cases a spectral learner would perform better if it could use some useful higher-order basis functions. An important direction for future work will be to determine how to recognize and take advantage of useful higher-order basis functions without losing the good generalization performance of a low-order approach.

REFERENCES

[1] D. Donoho and I. Johnstone, "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 1994.

[2] ——, "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 1995.

[3] A. Drake and D. Ventura, "A practical generalization of Fourier-based learning," in *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 185–192.

[4] ——, "Search techniques for Fourier-based learning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2009, pp. 1040–1045.

[5] J. Jackson, "An efficient membership-query algorithm for learning DNF with respect to the uniform distribution," *Journal of Computer and System Sciences*, vol. 55, pp. 414–440, 1997.

[6] H. Kargupta and B. Park, "A Fourier Spectrum-Based Approach to Represent Decision Trees for Mining Data Streams in Mobile Environments," *IEEE Transactions on Knowledge and Data Engineering*, 2004.

[7] H. Kargupta, B. Park, D. Hershberger, and E. Johnson, "Collective Data Mining: A New Perspective Toward Distributed Data Mining," in *Advances in Distributed and Parallel Knowledge Discovery*. AAAI/MIT Press, 2000.

[8] E. Kushilevitz and Y. Mansour, "Learning decision trees using the Fourier spectrum," *SIAM Journal on Computing*, vol. 22, no. 6, pp. 1331–1348, 1993.

[9] N. Linial, Y. Mansour, and N. Nisan, "Constant depth circuits, Fourier transform, and learnability," *Journal of the ACM*, vol. 40, no. 3, pp. 607–620, 1993.

[10] Y. Mansour and S. Sahar, "Implementation issues in the Fourier transform algorithm," *Machine Learning*, vol. 14, pp. 5–33, 2000.

[11] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich, "Weakly Learning DNF and Characterizing Statistical Query Learning Using Fourier Analysis," in *Proceedings of the ACM Symposium on Theory of Computing*, 1994.

[12] N. Bshouty and C. Tamon, "On the Fourier Spectrum of Monotone Functions," *Journal of the ACM*, 1996.

[13] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the 13th International Conference on Machine Learning*, vol. 55, 1996, pp. 148–156.

[14] D. Newman, S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," 1998.