Kernel Affine Projection Algorithms

Weifeng Liu and José C. Príncipe

Abstract

The combination of the famed kernel trick and affine projection algorithms (APA) yields powerful nonlinear extensions, named collectively here KAPA. This paper is a follow-up study of the recently introduced kernel least-mean-square algorithm (KLMS). KAPA inherits the simplicity and online nature of KLMS while reducing its gradient noise, boosting performance. More interestingly, it provides a unifying model for several neural network techniques, including kernel least-mean-square algorithms, kernel adaline, sliding-window kernel recursive-least-squares (KRLS) and regularization networks. Therefore, many insights can be gained into the basic relations among them and the trade-off between computation complexity and performance. Several simulations illustrate its wide applicability.

Index Terms

Affine projection algorithms, kernel methods.

I. Introduction

The solid mathematical foundation, wide and successful applications are making kernel methods very popular. By the famed kernel trick, many linear methods have been recast in high dimensional reproducing kernel Hilbert spaces (RKHS) to yield more powerful nonlinear extensions, including support vector machines [1], principal component analysis [2], recursive least squares [3], Hebbian algorithm [4], Adaline [5], etc.

More recently, a kernelized least-mean-square (KLMS) algorithm was proposed in [6], which implicitly creates a growing radial basis function network (RBF) with a learning strategy similar to resource-allocating networks (RAN) proposed by Platt [7]. As an improvement, kernelized affine projection algorithms (KAPA) are presented for the first time in this paper by reformulating the conventional affine projection algorithm (APA) [8] in general reproducing kernel Hilbert spaces (RKHS). The new algorithms are online, simple, significantly reduce the gradient noise compared with the KLMS and thus improve performance.

More interestingly, the KAPA reduces to the kernel least-mean-square (KLMS), sliding-window kernel recursive least squares (SW-KRLS), kernel adaline and regularization networks naturally in special cases. Thus it provides a

Manuscript received Sept. 26, 2007. The authors are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611 USA, phone: 352-392-2682, fax: 352-392-0044 (email: weifeng@cnel.ufl.edu, principe@cnel.ufl.edu).

unifying model for these existing methods and helps better understand the basic relations among them and the trade-off between complexity and performance. Moreover, it also advances our understanding on the resource-allocating networks. Exploiting the underlying linear structure of RKHS, a brief discussion on its well-posedness will be conducted.

The organization of the paper is as follows. In section II, the affine projection algorithms are briefly reviewed. Next in section III, the kernel trick is applied to formulate the nonlinear affine projection algorithms. Other related algorithms are reviewed as special cases of the KAPA in section IV. We detail the implementation of the KAPA in section V. Three experiments are studied in section VI to support our theory. Finally section VII summarizes the conclusions and future lines of research.

The notation used throughout the paper is summarized in Table I.

TABLE I
NOTATIONS

	description	examples
Scalars	Small *italic* letters	d
Vectors	Small **bold** letters	$\mathbf{w}$, $\boldsymbol{\omega}$, $\mathbf{a}$
Matrices	Capital **BOLD** letters	$\mathbf{U}$, $\mathbf{\Phi}$
Time or iteration	indices in parentheses	$\mathbf{u}(i)$, $d(i)$
Components of vectors or matrices	subscript indices	$\mathbf{a}_j(i)$, $\mathbf{G}_{i,j}$

II. A Review of the Affine Projection Algorithms

Let d be a zero-mean scalar-valued random variable and let $\mathbf{u}$ be a zero-mean $L \times 1$ random variable with a positive-definite covariance matrix $\mathbf{R_u} = E[\mathbf{u}\mathbf{u}^T]$. The cross-covariance vector of d and $\mathbf{u}$ is denoted by $\mathbf{r}_{d\mathbf{u}} = E[d\mathbf{u}]$. The weight vector $\mathbf{w}$ that solves

$$\min_{\mathbf{w}} E|d - \mathbf{w}^T\mathbf{u}|^2 \tag{1}$$

is given by $\mathbf{w}^o = \mathbf{R}_\mathbf{u}^{-1}\mathbf{r}_{d\mathbf{u}}$ [8].

Several methods that approximate $\mathbf{w}$ iteratively also exist. For example, the common gradient method

$$\mathbf{w}(0) = \text{initial guess;} \quad \mathbf{w}(i) = \mathbf{w}(i-1) + \eta[\mathbf{r}_{d\mathbf{u}} - \mathbf{R_u}\mathbf{w}(i-1)] \tag{2}$$

or the regularized Newton's recursion,

$$\mathbf{w}(0) = \text{initial guess;} \quad \mathbf{w}(i) = \mathbf{w}(i-1) + \eta(\mathbf{R_u} + \varepsilon\mathbf{I})^{-1}[\mathbf{r}_{d\mathbf{u}} - \mathbf{R_u}\mathbf{w}(i-1)] \tag{3}$$

where ε is a small positive regularization factor and η is the step size specified by designers.

Stochastic-gradient algorithms replace the covariance matrix and the cross-covariance vector by local approximations directly from data at each iteration. There are several ways for obtaining such approximations. The trade-off is computation complexity, convergence performance, and steady-state behavior [8].

Assume that we have access to observations of the random variables d and $\mathbf{u}$ over time

$$\{d(1), d(2), \dots\} \quad \text{and} \quad \{\mathbf{u}(1), \mathbf{u}(2), \dots\}$$

The Least-mean-square (LMS) algorithm simply uses the instantaneous values for approximations $\hat{\mathbf{R}}_{\mathbf{u}} = \mathbf{u}(i)\mathbf{u}(i)^T$ and $\hat{\mathbf{r}}_{d\mathbf{u}} = d(i)\mathbf{u}(i)$. The corresponding steepest-descent recursion (2) and Newton's recursion (3) become

$$\mathbf{w}(i) = \mathbf{w}(i-1) + \eta\mathbf{u}(i)[d(i) - \mathbf{u}(i)^T\mathbf{w}(i-1)] \tag{4}$$

$$\mathbf{w}(i) = \mathbf{w}(i-1) + \eta\mathbf{u}(i)[\mathbf{u}(i)^T\mathbf{u}(i) + \varepsilon\mathbf{I}]^{-1}[d(i) - \mathbf{u}(i)^T\mathbf{w}(i-1)] \tag{5}$$

The affine projection algorithm however employs better approximations. Specifically, $\mathbf{R}_{\mathbf{u}}$ and $\mathbf{r}_{d\mathbf{u}}$ are replaced by the instantaneous approximations from the K most recent regressors and observations. Denoting

$$\mathbf{U}(i) = [\mathbf{u}(i-K+1), ..., \mathbf{u}(i)]_{L\times K} \quad \text{and} \quad \mathbf{d}(i) = [d(i-K+1), \dots, d(i)]^T$$

one has

$$\begin{aligned} \hat{\mathbf{R}}_{\mathbf{u}} &= \frac{1}{K}\mathbf{U}(i)\mathbf{U}(i)^T \\ \hat{\mathbf{r}}_{d\mathbf{u}} &= \frac{1}{K}\mathbf{U}(i)\mathbf{d}(i) \end{aligned} \tag{6}$$

Therefore (2) and (3) become

$$\mathbf{w}(i) = \mathbf{w}(i-1) + \eta\mathbf{U}(i)[\mathbf{d}(i) - \mathbf{U}(i)^T\mathbf{w}(i-1)] \tag{7}$$

$$\mathbf{w}(i) = \mathbf{w}(i-1) + \eta[\mathbf{U}(i)\mathbf{U}(i)^T + \varepsilon\mathbf{I}]^{-1}\mathbf{U}(i)[\mathbf{d}(i) - \mathbf{U}(i)^T\mathbf{w}(i-1)] \tag{8}$$

and (8), by the matrix inversion lemma, is equivalent to [8]

$$\mathbf{w}(i) = \mathbf{w}(i-1) + \eta\mathbf{U}(i)[\mathbf{U}(i)^T\mathbf{U}(i) + \varepsilon\mathbf{I}]^{-1}[\mathbf{d}(i) - \mathbf{U}(i)^T\mathbf{w}(i-1)] \tag{9}$$

It is noted that this equivalence lets us deal with the matrix $[\mathbf{U}(i)^T\mathbf{U}(i) + \varepsilon\mathbf{I}]$ instead of $[\mathbf{U}(i)\mathbf{U}(i)^T + \varepsilon\mathbf{I}]$ and it plays a very important role in the derivation of kernel extensions. We call recursion (7) APA-1 and recursion (9) APA-2.

In some circumstances, a regularized solution is needed instead of (1). The regularized LS problem is

$$\min_{\mathbf{w}} E|d - \mathbf{w}^T\mathbf{u}|^2 + \lambda||\mathbf{w}||^2 \tag{10}$$

where λ is the regularization parameter (not the regularization factor ε in Newton's recursion). The gradient method is

$$\begin{aligned}\mathbf{w}(i) &= \mathbf{w}(i-1) + \eta[\mathbf{r}_{d\mathbf{u}} - (\lambda\mathbf{I} + \mathbf{R}_{\mathbf{u}})w(i-1)] \\ &= (1-\eta\lambda)w(i-1) + \eta[\mathbf{r}_{d\mathbf{u}} - \mathbf{R}_{\mathbf{u}}w(i-1)]\end{aligned} \tag{11}$$

The Newton's recursion with $\varepsilon = 0$ is

$$\begin{aligned}\mathbf{w}(i) &= \mathbf{w}(i-1) + \eta(\lambda\mathbf{I} + \mathbf{R}_{\mathbf{u}})^{-1}[\mathbf{r}_{d\mathbf{u}} - (\lambda\mathbf{I} + \mathbf{R}_{\mathbf{u}})\mathbf{w}(i-1)] \\ &= (1-\eta)\mathbf{w}(i-1) + \eta(\lambda\mathbf{I} + \mathbf{R}_{\mathbf{u}})^{-1}\mathbf{r}_{d\mathbf{u}}\end{aligned} \tag{12}$$

If the approximations (6) are used, we have

$$\mathbf{w}(i) = (1-\eta\lambda)\mathbf{w}(i-1) + \eta\mathbf{U}(i)[\mathbf{d}(i) - \mathbf{U}(i)^T\mathbf{w}(i-1)] \tag{13}$$

and

$$\mathbf{w}(i) = (1-\eta)\mathbf{w}(i-1) + \eta[\lambda\mathbf{I} + \mathbf{U}(i)\mathbf{U}(i)^T]^{-1}\mathbf{U}(i)\mathbf{d}(i) \tag{14}$$

which is, by the matrix inversion lemma, equivalent to

$$\mathbf{w}(i) = (1-\eta)\mathbf{w}(i-1) + \eta\mathbf{U}(i)[\lambda\mathbf{I} + \mathbf{U}(i)^T\mathbf{U}(i)]^{-1}\mathbf{d}(i) \tag{15}$$

For simplicity, recursions (13) and (15) are named here APA-3 and APA-4 respectively.

III. The Kernel Affine Projection Algorithms

A kernel [9] is a continuous, symmetric, positive-definite function $\kappa : \mathbb{U} \times \mathbb{U} \to \mathbb{R}$. $\mathbb{U}$ is the input domain, a compact subset of $\mathbb{R}^L$. The commonly used kernels include the Gaussian kernel (16) and the polynomial kernel (17):

$$\kappa(\mathbf{u}, \mathbf{u}') = \exp(-a||\mathbf{u} - \mathbf{u}'||^2) \tag{16}$$

$$\kappa(\mathbf{u}, \mathbf{u}') = (\mathbf{u}^T\mathbf{u}' + 1)^p \tag{17}$$

The Mercer theorem [9], [10] states that any kernel $\kappa(\mathbf{u}, \mathbf{u}')$ can be expanded as follows:

$$\kappa(\mathbf{u}, \mathbf{u}') = \sum_{i=1}^{\infty} \varsigma_i\phi_i(\mathbf{u})\phi_i(\mathbf{u}') \tag{18}$$

where ς_i and ϕ_i are the eigenvalues and the eigenfunctions respectively. The eigenvalues are non-negative. Therefore, a mapping φ can be constructed as

$$\varphi : \mathbb{U} \rightarrow \mathbb{F}$$
$$\varphi(\mathbf{u}) = [\sqrt{\varsigma_1}\phi_1(\mathbf{u}), \sqrt{\varsigma_2}\phi_2(\mathbf{u}), ...]$$

such that

$$\kappa(\mathbf{u}, \mathbf{u}') = \varphi(\mathbf{u})^T \varphi(\mathbf{u}') \tag{19}$$

By construction, the dimensionality of $\mathbb{F}$ is determined by the number of strictly positive eigenvalues, which can be infinite in the Gaussian kernel case.

We utilize this theorem to transform the data $\mathbf{u}(i)$ into the feature space $\mathbb{F}$ as $\varphi(\mathbf{u}(i))$ and interpret (19) as the usual dot product. Denoting $\varphi(i) = \varphi(\mathbf{u}(i))$, we formulate the affine projection algorithms on the example sequence $\{d(1), d(2), \dots\}$ and $\{\varphi(1), \varphi(2), \dots\}$ to estimate the weight vector $\boldsymbol{\omega}$ that solves

$$\min_{\boldsymbol{\omega}} E|d - \boldsymbol{\omega}^T \varphi(\mathbf{u})|^2 \tag{20}$$

By straightforward manipulation, (7) becomes

$$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta \boldsymbol{\Phi}(i)[\mathbf{d}(i) - \boldsymbol{\Phi}(i)^T \boldsymbol{\omega}(i-1)] \tag{21}$$

and (9) becomes

$$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta \boldsymbol{\Phi}(i)[\boldsymbol{\Phi}(i)^T \boldsymbol{\Phi}(i) + \varepsilon \mathbf{I}]^{-1}[\mathbf{d}(i) - \boldsymbol{\Phi}(i)^T \boldsymbol{\omega}(i-1)] \tag{22}$$

where $\boldsymbol{\Phi}(i) = [\varphi(i-K+1), ..., \varphi(i)]$.

Accordingly, (13) becomes

$$\boldsymbol{\omega}(i) = (1 - \lambda\eta)\boldsymbol{\omega}(i-1) + \eta \boldsymbol{\Phi}(i)[\mathbf{d}(i) - \boldsymbol{\Phi}(i)^T \boldsymbol{\omega}(i-1)] \tag{23}$$

and (15) becomes

$$\boldsymbol{\omega}(i) = (1 - \eta)\boldsymbol{\omega}(i-1) + \eta \boldsymbol{\Phi}(i)[\boldsymbol{\Phi}(i)^T \boldsymbol{\Phi}(i) + \lambda \mathbf{I}]^{-1}\mathbf{d}(i) \tag{24}$$

For simplicity, we refer to the recursions (21), (22), (23), and (24) as KAPA-1, KAPA-2, KAPA-3, and KAPA-4 respectively.

A. Kernel Affine Projection Algorithm (KAPA-1)

It may be difficult to have direct access to the weights and the transformed data in feature space, so (21) needs to be modified. If we set the initial guess $\boldsymbol{\omega}(0) = 0$, the iteration of (21) will be

$$
\begin{aligned}
&\boldsymbol{\omega}(0) = 0, \\
&\boldsymbol{\omega}(1) = \eta d(1)\varphi(1) = \mathbf{a}_1(1)\varphi(1), \\
&... \\
&\boldsymbol{\omega}(i-1) = \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\varphi(j), \\
&\boldsymbol{\Phi}(i)^T\boldsymbol{\omega}(i-1) = [\sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\kappa_{i-K+1,j}, , ..., \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\kappa_{i-1,j}, \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\kappa_{i,j}]^T, \\
&\mathbf{e}(i) = \mathbf{d}(i) - \boldsymbol{\Phi}(i)^T\boldsymbol{\omega}(i-1), \\
&\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}(i)\mathbf{e}(i) = \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\varphi(j) + \sum_{j=1}^{K} \eta\mathbf{e}_j(i)\varphi(i-j+K).
\end{aligned}
\tag{25}
$$

where $\kappa_{i,j} = \kappa(\mathbf{u}(i), \mathbf{u}(j))$ for simplicity.

Note that during the iteration, the weight vector in the feature space assumes the following expansion

$$
\boldsymbol{\omega}(i) = \sum_{j=1}^{i} \mathbf{a}_j(i)\varphi(j) \quad \forall\, i > 0 \tag{26}
$$

i.e. the weight at time i is a linear combination of the previous transformed input. This result may seem simply a restatement of the representer theorem in [11]. However, it should be emphasized that this result does not rely on any explicit minimal norm constraint as required for the representer theorem. As pointed out in [12], the gradient search involved has an inherent regularization mechanism which guarantees the solution is in the data subspace under appropriate initialization. In general, the initialization $\boldsymbol{\omega}(0)$ can introduce whatever apriori information is available, which can be any linear combination of any transformed data in order to utilize the kernel trick.

By (26), the updating on the weight vector reduces to the updating on the expansion coefficients

$$
\mathbf{a}_k(i) = \begin{cases}
\eta(d(i) - \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\kappa_{i,j}), & k = i \\
\mathbf{a}_k(i-1) + \eta(d(k) - \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\kappa_{k,j}), & i-K+1 \le k \le i-1 \\
\mathbf{a}_k(i-1), & 1 \le k < i-K+1
\end{cases} \tag{27}
$$

Since $\mathbf{e}_{i+1-k}(i) = d(k) - \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\kappa_{k,j}$ is the prediction error of data $\{\mathbf{u}(k), d(k)\}$ by the network $\boldsymbol{\omega}(i-1)$, the interpretation of (27) is straightforward: allocate a new unit with coefficient $\eta\mathbf{e}_1(i)$ and update the coefficients for the other $K-1$ most recent units by $\eta\mathbf{e}_{i+1-k}(i)$ for $i-K+1 \le k \le i-1$.

The pseudocode for KAPA-1 is listed in Algorithm 1.

Algorithm 1 Kernel Affine Projection Algorithm (KAPA-1)

Initialization:
learning step η
$\mathbf{a}_1(1) = \eta d(1)$
while $\{\mathbf{u}(i), d(i)\}$ available **do**
  %allocate a new unit
  $\mathbf{a}_i(i-1) = 0$
  for $k = \max(1, i-K+1)$ to i **do**
    %evaluate outputs of the current network
    $y(i,k) = \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\kappa_{k,j}$
    %computer errors
    $e(i,k) = d(k) - y(i,k)$
    %update the $\min(i, K)$ most recent units
    $\mathbf{a}_k(i) = \mathbf{a}_k(i-1) + \eta e(i,k)$
  end for
  if $i > K$ **then**
    %keep the remaining
    for $k = 1$ to $i-K$ **do**
      $\mathbf{a}_k(i) = \mathbf{a}_k(i-1)$
    end for
  end if
end while

B. Normalized KAPA (KAPA-2)

Similarly, the regularized Newton's recursion (22) can be factorized into the following steps

$$\begin{aligned}
\boldsymbol{\omega}(i-1) &= \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\varphi(j), \\
\mathbf{e}(i) &= \mathbf{d}(i) - \boldsymbol{\Phi}(i)^T \boldsymbol{\omega}(i-1), \\
\mathbf{G}(i) &= \boldsymbol{\Phi}(i)^T \boldsymbol{\Phi}(i), \\
\boldsymbol{\omega}(i) &= \boldsymbol{\omega}(i-1) + \eta \boldsymbol{\Phi}(i)[\mathbf{G}(i) + \varepsilon \mathbf{I}]^{-1} \mathbf{e}(i).
\end{aligned} \tag{28}$$

In practice, we do not have access to the transformed weight $\boldsymbol{\omega}$ or any transformed data, so the update has to be on the expansion coefficient $\mathbf{a}$ like in KAPA-1. The whole recursion is similar to the KAPA-1 except that the error is normalized by a $K \times K$ matrix $[\mathbf{G}(i) + \varepsilon \mathbf{I}]^{-1}$.

C. Leaky KAPA (KAPA-3)

The feature space may be infinite dimensional depending on the chosen kernel, which may cause the cost function (20) to be ill-posed in the conventional empirical risk minimization (ERM) sense [13]. The common practice is to constrain the solution norm:

$$\min_{\boldsymbol{\omega}} E|d - \boldsymbol{\omega}^T \varphi(\mathbf{u})|^2 + \lambda ||\boldsymbol{\omega}||^2 \tag{29}$$

TABLE II
COMPARISON OF FOUR KAPA UPDATE RULES

Algorithm	Update equation
KAPA-1	$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}(i)[\mathbf{d}(i) - \boldsymbol{\Phi}(i)^T\boldsymbol{\omega}(i-1)]$
KAPA-2	$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}(i)[\boldsymbol{\Phi}(i)^T\boldsymbol{\Phi}(i) + \varepsilon\mathbf{I}]^{-1}[\mathbf{d}(i) - \boldsymbol{\Phi}(i)^T\boldsymbol{\omega}(i-1)]$
KAPA-3	$\boldsymbol{\omega}(i) = (1-\lambda\eta)\boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}(i)[\mathbf{d}(i) - \boldsymbol{\Phi}(i)^T\boldsymbol{\omega}(i-1)]$
KAPA-4	$\boldsymbol{\omega}(i) = (1-\eta)\boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}(i)[\boldsymbol{\Phi}(i)^T\boldsymbol{\Phi}(i) + \lambda\mathbf{I}]^{-1}\mathbf{d}(i)$

As we have already shown in (23), the leaky KAPA is

$$\boldsymbol{\omega}(i) = (1-\lambda\eta)\boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}(i)[\mathbf{d}(i) - \boldsymbol{\Phi}(i)^T\boldsymbol{\omega}(i-1)] \tag{30}$$

Again, the iteration will be on the expansion coefficient $\mathbf{a}$, which is similar to the KAPA-1.

$$\mathbf{a}_k(i) = \begin{cases} \eta(d(i) - \sum_{j=1}^{i-1}\mathbf{a}_j(i-1)\kappa_{i,j}), & k = i \\ (1-\lambda\eta)\mathbf{a}_k(i-1) + \eta(d(k) - \sum_{j=1}^{i-1}\mathbf{a}_j(i-1)\kappa_{k,j}), & i-K+1 \le k \le i-1 \\ (1-\lambda\eta)\mathbf{a}_k(i-1), & 1 \le k < i-K+1 \end{cases} \tag{31}$$

The only difference is that KAPA-3 has a scaling factor $(1-\lambda\eta)$ multiplying the previous weight, which is usually less than 1, and it imposes a forgetting mechanism so that the training data in the far past are scaled down exponentially. Furthermore since the network size is growing over training a transformed data can be pruned from the expansion easily if its coefficient is smaller than some pre-specified threshold. For large data sets, the growing nature of these family of algorithms poses a big problem for implementations, therefore network size control is very important. We will discuss it more in the sparsification section.

D. Leaky KAPA with Newton's Recursion (KAPA-4)

As before, the KAPA-4 (24) reduces to

$$\mathbf{a}_k(i) = \begin{cases} \eta d(i), & k = i \\ (1-\eta)\mathbf{a}_k(i-1) + \eta d(k), & i-K+1 \le k \le i-1 \\ (1-\eta)\mathbf{a}_k(i-1), & 1 \le k < i-K+1 \end{cases} \tag{32}$$

Among these four algorithms, the first three require the error information to update the network which is computationally expensive. Therefore the different update rule in KAPA-4 has a huge significance in terms of computation since it only needs a $K \times K$ matrix inversion, which by using the sliding-window trick only requires $O(K^2)$ operations [14].

We summarize the four KAPA update equations in Table II for convenience.

IV. A TAXONOMY FOR RELATED ALGORITHMS

A. Kernel Least-Mean-Square Algorithm (KAPA-1, $K = 1$)

If $K = 1$, KAPA-1 reduces to the following kernel least-mean-square algorithm (KLMS) introduced in [6]

$$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta\varphi(i)[d(i) - \varphi(i)^T\boldsymbol{\omega}(i-1)]$$

It is not difficult to verify that the weight vector assumes the following expansion

$$\boldsymbol{\omega}(i) = \sum_{j=1}^{i} e(j)\varphi(j)$$

where $e(j) = d(j) - \boldsymbol{\omega}(j-1)^T\varphi(j)$ is the apriori error.

It is seen that the KLMS allocates a new unit when a new training data comes in with the input $\mathbf{u}(i)$ as the center and the prediction error as the coefficient (scaled by the step size). In other words, once the unit is allocated, the coefficient is fixed. It mimics the resource-allocating step in the RAN algorithm whereas it neglects the adaptation step. In this sense, the KAPA algorithms that allocate a new unit for the present input and also adapt the other $K-1$ most recent allocated units, are closer to the original RAN.

The normalized version of the KLMS is as follows (NKLMS):

$$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \frac{\eta\varphi(i)}{\varepsilon + \kappa_{i,i}}[d(i) - \varphi(i)^T\boldsymbol{\omega}(i-1)] \tag{33}$$

Notice that for translation invariant kernels, i.e., $\kappa_{i,i} = const$, the KLMS is automatically normalized. Sometimes we use KLMS-1 and KLMS-2 to distinguish the two.

B. Norma (KAPA-3, $K = 1$)

Similarly the KAPA-3 (23) reduces to the Norma algorithm introduced by Kivinen in [15].

$$\boldsymbol{\omega}(i) = (1-\eta\lambda)\boldsymbol{\omega}(i-1) + \eta\varphi(i)[d(i) - \varphi(i)^T\boldsymbol{\omega}(i-1)] \tag{34}$$

C. Kernel Adaline (KAPA-1, $K = N$)

Assume that the size of the training data is finite N. If we set $K = N$, then the update rule of the KAPA-1 becomes

$$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}[\mathbf{d} - \boldsymbol{\Phi}^T\boldsymbol{\omega}(i-1)]$$

where the full data matrices are

$$\boldsymbol{\Phi} = [\varphi(1), ..., \varphi(N)], \quad \mathbf{d} = [d(1), ..., d(N)]$$

It is easy to check that the weight vector also assumes the following expansion

$$\boldsymbol{\omega}(i) = \sum_{j=1}^{N} \mathbf{a}_j(i)\varphi(j)$$

And the updating on the expansion coefficients is

$$\mathbf{a}_j(i) = \mathbf{a}_j(i-1) + \eta[d(j) - \varphi(j)^T\boldsymbol{\omega}(i-1)]$$

This is nothing but the kernel adaline introduced in [5]. Notice the fact that the kernel adaline is not an online method.

D. Recursively-adapted Radial Basis Function Networks (KAPA-3, $\eta\lambda = 1$, $K = N$)

Assume the size of the training data is N as above. If we set $\eta\lambda = 1$ and $K = N$, the update rule of KAPA-3 becomes

$$\boldsymbol{\omega}(i) = \eta\boldsymbol{\Phi}[\mathbf{d} - \boldsymbol{\Phi}^T\boldsymbol{\omega}(i-1)]$$

which is the recursively-adapted RBF (RA-RBF) network introduced in [16]. This is a very intriguing algorithm using the 'global' error directly to compose the new network. By contrast, the KLMS-1 uses the apriori errors to compose the network.

E. Sliding-window Kernel RLS (KAPA-4, $\eta = 1$)

In KAPA-4, if we set $\eta = 1$, we have

$$\boldsymbol{\omega}(i) = \boldsymbol{\Phi}(i)[\boldsymbol{\Phi}(i)^T\boldsymbol{\Phi}(i) + \lambda\mathbf{I}]^{-1}\mathbf{d}(i) \tag{35}$$

which is the sliding-window kernel RLS (SW-KRLS) introduced in [14]. The inverse operation of the sliding-window Gram matrix can be simplified to $O(K^2)$.

F. Regularization Networks (KAPA-4, $\eta = 1$, $K = N$)

We assume there are only N training data and $K = N$. Equation (24) becomes directly

$$\boldsymbol{\omega}(i) = \boldsymbol{\Phi}[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I}]^{-1}\mathbf{d} \tag{36}$$

which is the regularization network (RegNet) [13].

We summarize all the related algorithms in Table III for convenience.

V. KAPA Implementation

In this section, we will discuss the implementation of the KAPA algorithms in detail.

TABLE III
LIST OF RELATED ALGORITHMS

Algorithm	Update equation	Relation to KAPA
KLMS	$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta\varphi(i)[d(i) - \varphi(i)^T\boldsymbol{\omega}(i-1)]$	KAPA-1, $K = 1$
NKLMS	$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \frac{\eta\varphi(i)}{(\varepsilon+\kappa_{i,i})}[d(i) - \varphi(i)^T\boldsymbol{\omega}(i-1)]$	KAPA-2, $K = 1$
Norma	$\boldsymbol{\omega}(i) = (1-\eta\lambda)\boldsymbol{\omega}(i-1) + \eta\varphi(i)[d(i) - \varphi(i)^T\boldsymbol{\omega}(i-1)]$	KAPA-3, $K = 1$
Kernel Adaline	$\boldsymbol{\omega}(i) = \boldsymbol{\omega}(i-1) + \eta\boldsymbol{\Phi}[\mathbf{d} - \boldsymbol{\Phi}^T\boldsymbol{\omega}(i-1)$	KAPA-1, $K = N$
RA-RBF	$\boldsymbol{\omega}(i) = \eta\boldsymbol{\Phi}[\mathbf{d} - \boldsymbol{\Phi}^T\boldsymbol{\omega}(i-1)]$	KAPA-3, $\eta\lambda = 1$, $K = N$
SW-KRLS	$\boldsymbol{\omega}(i) = \boldsymbol{\Phi}(i)[\boldsymbol{\Phi}(i)^T\boldsymbol{\Phi}(i) + \lambda\mathbf{I}]^{-1}\mathbf{d}(i)$	KAPA-4, $\eta = 1$
RegNet	$\boldsymbol{\omega}(i) = \boldsymbol{\Phi}[\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\mathbf{I}]^{-1}\mathbf{d}$	KAPA-4, $\eta = 1$, $K = N$

A. Error reusing

As we see in KAPA-1, KAPA-2 and KAPA-3, the most time-consuming part of the computation is to obtain the error information. For example, suppose $\boldsymbol{\omega}(i-1) = \sum_{j=1}^{i-1} \mathbf{a}_j(i-1)\varphi(j)$. We need to calculate $e(i,k) = d(k) - \boldsymbol{\omega}(i-1)^T\varphi(k)$ $(i-K+1 \le k \le i)$ to compute $\boldsymbol{\omega}(i)$, which consists of $(i-1)K$ kernel evaluations. As i increases, this dominates the computation time. In this sense, the computation complexity of the KAPA is K times of the KLMS. However, after a careful manipulation, we can shrink the complexity gap between KAPA and the KLMS.

Assume that we store all the K errors $e(i-1,k) = d(k) - \boldsymbol{\omega}(i-2)^T\varphi(k)$ for $i-K \le k \le i-1$ from the previous iteration. At the present iteration, we have

$$\begin{aligned} e(i,k) &= d(k) - \varphi(k)^T\boldsymbol{\omega}(i-1) \\ &= d(k) - \varphi(k)^T[\boldsymbol{\omega}(i-2) + \eta\sum_{j=i-K}^{i-1} e(i-1,j)\varphi(j)] \\ &= [d(k) - \varphi(k)^T\boldsymbol{\omega}(i-2)] + \eta\sum_{j=i-K}^{i-1} e(i-1,j)\kappa_{j,k} \\ &= e(i-1,k) + \sum_{j=i-K}^{i-1} \eta e(i-1,j)\kappa_{j,k} \end{aligned} \tag{37}$$

Since $e(i-1,i)$ has not been computed yet, we have to calculate $e(i,i)$ by $i-1$ times kernel evaluation anyway. Overall the computation complexity of the KAPA-1 is $O(i+K^2)$, which is $O(K^2)$ more than the KLMS.

B. Sliding-window Gram Matrix Inversion

In KAPA-2 and KAPA-4, another computation difficulty is to invert a $K \times K$ matrix, which normally requires $O(K^3)$. However, in the KAPA, the data matrix $\boldsymbol{\Phi}(i)$ has a sliding window structure, therefore a trick can be used to speed up the computation. The trick is based on the matrix inversion formula and was introduced in [14]. We

outline the basic calculation steps here. Suppose the sliding matrices share the same sub-matrix $\mathbf{D}$

$$\mathbf{G}(i-1)+\lambda\mathbf{I}=\begin{bmatrix} a & \mathbf{b}^T \\ \mathbf{b} & \mathbf{D} \end{bmatrix}, \quad \mathbf{G}(i)+\lambda\mathbf{I}=\begin{bmatrix} \mathbf{D} & \mathbf{h} \\ \mathbf{h}^T & g \end{bmatrix} \tag{38}$$

and we know from the previous iteration

$$(\mathbf{G}(i-1)+\lambda\mathbf{I})^{-1}=\begin{bmatrix} e & \mathbf{f}^T \\ \mathbf{f} & \mathbf{H} \end{bmatrix} \tag{39}$$

First we need to calculate the inverse of $\mathbf{D}$ as

$$\mathbf{D}^{-1}=\mathbf{H}-\mathbf{f}\mathbf{f}^T/e \tag{40}$$

Then we can update the inverse of the new Gram matrix as

$$(\mathbf{G}(i)+\lambda\mathbf{I})^{-1}=\begin{bmatrix} \mathbf{D}^{-1}+(\mathbf{D}^{-1}\mathbf{h})(\mathbf{D}^{-1}\mathbf{h})^T s & -(\mathbf{D}^{-1}\mathbf{h})s \\ -(\mathbf{D}^{-1}\mathbf{h})^T s & s \end{bmatrix} \tag{41}$$

with $s=(g-\mathbf{h}^T\mathbf{D}^{-1}\mathbf{h})^{-1}$. s^{-1} is the Schur complement of $\mathbf{D}$ in $(\mathbf{G}(i)+\lambda\mathbf{I})$, which actually measures the distance of the new data to the other $K-1$ most recent data in the feature space. The overall complexity is $O(K^2)$.

C. Sparsification

A sparse model is desired because it reduces the complexity in terms of computation and memory, and it usually yields better generalization [3]. On the other hand, in the context of adaptive filtering, training data may just be available sequentially, i.e., one at a time. As we see in the formulation of KAPA, the network size increases linearly with the number of training data, which may pose a big problem for the KAPA algorithms to be applied in online applications. The sparse model idea is inspired by Vapnik's support vector machines. It is also introduced in [7] with the novelty criterion and extensively studied in [3] under approximate linear dependency (ALD). There are many other ways to achieve sparseness that require the creation of a basis dictionary and storage of the corresponding coefficients. Suppose the present dictionary is $\mathcal{D}(i)=\{\mathbf{c}_j\}_{j=1}^{m(i)}$ where $\mathbf{c}_j$ is the jth center and $m(i)$ is the cardinality. When a new data pair $\{\mathbf{u}(i+1), d(i+1)\}$ is presented, a decision is made immediately whether $\mathbf{u}(i+1)$ should be added into the dictionary as a center.

The novelty criterion introduced by Platt is relatively simple. First it calculates the distance of $\mathbf{u}(i+1)$ to the present dictionary $dis_1=\min_{\mathbf{c}_j\in\mathcal{D}(i)}||\mathbf{u}(i+1)-\mathbf{c}_j||$. If it is smaller than some preset threshold, say δ_1, $\mathbf{u}(i+1)$ will not be added into the dictionary. Otherwise, the method computes the prediction error $e(i+1,i+1)=d(i+1)-\varphi(i+1)^T\boldsymbol{\omega}(i)$. Only if the prediction error is larger than another preset threshold, say δ_2, $\mathbf{u}(i+1)$ will be accepted as a new center.

The ALD test introduced in [3] is more computationally involved. It tests the following cost $dis_2 = \min_{\forall \mathbf{b}} ||\varphi(\mathbf{u}(i+1)) - \sum_{\mathbf{c}_j \in \mathcal{D}(i)} \mathbf{b}_j \varphi(\mathbf{c}_j)||$ which indicates the distance of the new input to the linear span of the present dictionary in the feature space. It turns out that dis_2 is the Schur complement of the Gram matrix of the present dictionary. As we saw in the previous section, this result can be used to get the new Gram matrix inverse if $\mathbf{u}(i+1)$ is accepted into the dictionary. Therefore this method is more suitable for the KAPA-2 and KAPA-4 because of efficiency. This link is very interesting since it reveals that the ALD test actually guarantees the invertibility of the new Gram matrix.

In the sparse model, if the new data is determined to be 'novel', the $K-1$ most recent data points in the dictionary is used to form the data matrix $\mathbf{\Phi}(i)$ together with the new data. Therefore a new unit is allocated and the update is on the $K-1$ most recent units in the dictionary. If the new data is determined to be not 'novel', it is simply discarded in this paper but different strategy can be employed to utilize the information like in [7] and [3].

The important consequences of the sparsification procedure are as follows:

1) If the input domain $\mathbb{U}$ is a compact set, the cardinality of the dictionary is always finite and upper bounded. This statement is not hard to prove using the finite covering theorem of the compact set and the fact that elements in the dictionary are δ-separable [3]. Here is the brief idea: suppose spheres with diameter δ are used to cover $\mathbb{U}$ and the optimal covering number is N. Then because any two centers in the dictionary can not be in the same sphere, the total number of the centers will be no greater than N regardless of the distribution and temporal structure of $\mathbf{u}$. Of course this is a worst case upper bound. In the case of finite training data, the network size will be finite anyway. This is true in applications like channel equalization, where the training sequence is part of each transmission frame. In a stationary environment, the network converges quickly and the threshold on prediction errors plays its part to constrain the network size. We will validate this claim in the simulation section. In a non-stationary environment, more sophisticated pruning methods should be used to constrain the network size. Simple strategies include pruning the oldest unit in the dictionary [14], pruning randomly [17], pruning the unit with the least coefficient or similar [18], [19]. It should be pointed out that the scalability issue is at the core of the kernel methods and so all the kernel methods need to deal with it in one way or the other. Indeed, the sequential nature of the KAPA enables active learning [20], [21] on huge data sets which is impossible in batch mode algorithms like regularization networks. The discussion on active learning with the KAPA is out of the scope of this paper and will be part of the future work.

2) Based on 1), we can prove that the solution norms of KLMS-1 and KAPA-1 are upper bounded [12].

The significance of 1) is of practical interest because it states that the system complexity is controlled by the novelty criterion parameters and designers can estimate a worst case upper bound. The significance of 2) is of theoretical interest because it guarantees the well-posedness of the algorithms. The well-posedness of the KAPA-3

TABLE IV
PERFORMANCE COMPARISON IN MG TIME SERIES PREDICTION

Algorithm	Parameters	Test Mean Square Error
LMS	$\eta = 0.04$	0.0208 ± 0.0009
KLMS	$\eta = 0.02$	0.0052 ± 0.00022
SW-KRLS	$K = 50$, $\lambda = 0.1$	0.0052 ± 0.00026
KAPA-1	$\eta = 0.03$, $K = 10$	0.0048 ± 0.00023
KAPA-2	$\eta = 0.03$, $K = 10$, $\epsilon = 0.1$	0.0040 ± 0.00028
KRLS	$\lambda = 0.1$	0.0027 ± 0.00009

and KAPA-4 is mostly ensured by the regularization term. See [13] and [14] for details.

VI. SIMULATIONS

A. Time Series Prediction

The first example is the short-term prediction of the Mackey-Glass (MG) chaotic time series [22], [23]. It is generated from the following time delay ordinary differential equation

$$\frac{dx(t)}{dt} = -bx(t) + \frac{ax(t-\tau)}{1 + x(t-\tau)^{10}} \tag{42}$$

with $b = 0.1$, $a = 0.2$, and $\tau = 30$. The time series is discretized at a sampling period of 6 seconds. The time embedding is 7, i.e. $\mathbf{u}(i) = [x(i-7), x(i-6), ..., x(i-1)]^T$ are used as the input to predict the present one $x(i)$ which is the desired response here. A segment of 500 samples is used as the training data and another 100 points as the test data (in the testing phase, the filter is fixed). All the data is corrupted by Gaussian noise with zero mean and 0.001 variance.

We compare the prediction performance of KLMS, KAPA-1, KAPA-2, KRLS, and a linear combiner trained with LMS. A Gaussian kernel with kernel parameter $a = 1$ in (16) is chosen for all the kernel-based algorithms. One hundred Monte Carlo simulations are run with different realizations of noise. The results are summarized in Tables IV. Fig. 1 is the learning curves for the LMS, KLMS-1, KAPA-1, KAPA-2 ($K = 10$) and KRLS respectively. As expected, the KAPA outperforms the KLMS.

As we can see in Table IV, the performance of the KAPA-2 is substantially better than the KLMS. All the results in the tables are in the form of 'average $\pm$ standard deviation'. Table V summarizes the computational complexity of these algorithms. The KLMS and KAPA effectively reduce the computational complexity and memory storage when compared with the KRLS. KAPA-3 and sliding-window KRLS are also tested on this problem. It is observed that the performance of the KAPA-3 is similar to KAPA-1 when the forgetting term is very close to 1 as expected and the results are severely biased when the forgetting term is reduced further. The reason can be found in [12].

TABLE V
COMPLEXITY COMPARISON AT ITERATION i

Algorithm	Computation	Memory
LMS	$O(L)$	$O(L)$
KLMS	$O(i)$	$O(i)$
SW-KRLS	$O(K^2)$	$O(K^2)$
KAPA-1	$O(i+K^2)$	$O(i+K)$
KAPA-2	$O(i+K^2)$	$O(i+K^2)$
KAPA-4	$O(K^2)$	$O(i+K^2)$
KRLS	$O(i^2)$	$O(i^2)$

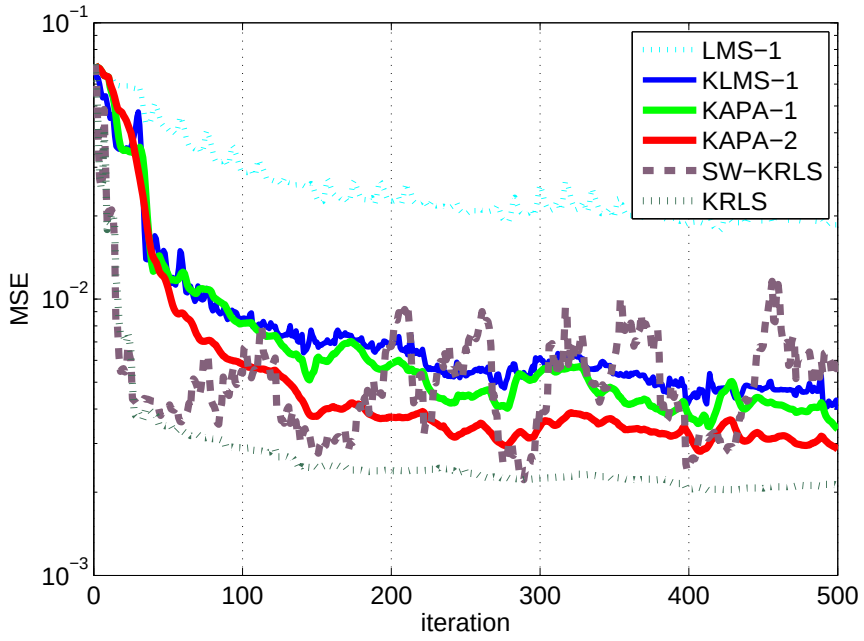

Fig. 1. The learning curves of the LMS, KLMS, KAPA-1 ($K = 10$), KAPA-2 ($K = 10$), SW-KRLS ($K = 50$) and KRLS

The performance of the sliding-window KRLS is included in the Fig. 1 and in Table IV with $K = 50$. It is observed that KAPA-4 (including the sliding-window KRLS) does not perform well with small K (< 50).

Next, we test how the novelty criterion affects the performance. A segment of 1000 samples is used as the training data and another 100 as the test data. All the data is corrupted by Gaussian noise with zero mean and 0.001 variance. The thresholds in the novelty criterion are set as $\delta_1 = 0.02$ and $\delta_2 = 0.06$. The learning curves are shown in Fig. 2 and the results are summarized in Table VI. It is seen that the complexity can be reduced dramatically with the novelty criterion with slight performance degeneration. Here SKLMS and SKAPA denote the sparse KLMS and the sparse KAPA respectively.

Several comments follow: Although formally being adaptive filters, these algorithms can be viewed as efficient alternatives to batch mode RBF networks, therefore it is practical to freeze their weights during test phase. Moreover,

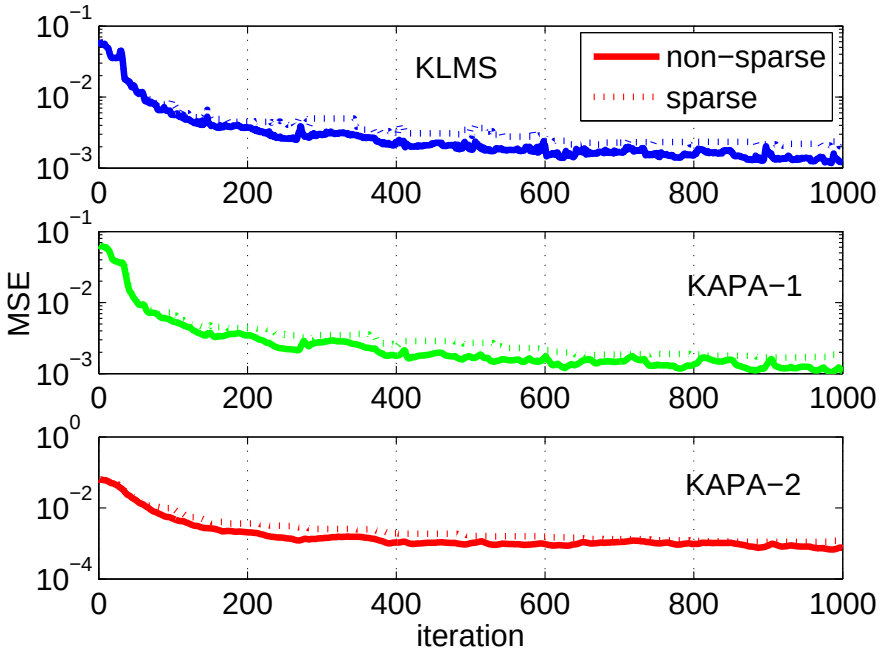

Fig. 2. The learning curves of the KLMS-1, KAPA-1 ($K = 10$) and KAPA-2 ($K = 10$) with and without sparsification

TABLE VI
PERFORMANCE COMPARISON IN MG TIME SERIES PREDICTION ON NOVELTY CRITERION

Algorithm	Parameters	Test Mean Square Error	Dictionary size
KLMS-1	$\eta = 0.02$	0.0015 ± 0.00012	1000
SKLMS-1	$\eta = 0.02$	0.0021 ± 0.00017	220
KAPA-1	$\eta = 0.03$	0.0012 ± 0.00014	1000
SKAPA-1	$\eta = 0.03$	0.0017 ± 0.00016	209
KAPA-2	$\eta = 0.03$, $\epsilon = 0.1$	0.0007 ± 0.00010	1000
SKAPA-2	$\eta = 0.03$, $\epsilon = 0.1$	0.0011 ± 0.00016	195

when compared with other nonlinear filters such as RBFs, we divide the data in training and testing as normally done in neural networks. Of course, it is also feasible to use the apriori prediction error as a performance indicator like in conventional adaptive filtering literature.

B. Noise Cancellation

Another important problem in signal processing is noise cancellation in which an unknown interference has to be removed based on some reference measurement. The basic structure of a noise cancellation system is shown in Fig. 3. The primary signal is $s(i)$ and its noisy measurement $d(i)$ acts as the desired signal of the system. $n(i)$ is a white noise process which is unknown, and $u(i)$ is its reference measurement, i.e. a distorted version of the noise process through some distortion function, which is unknown in general. Here $u(i)$ is the input of the adaptive filter. The objective is to use $u(i)$ as the input to the filter and to obtain as the filter output an estimate of the noise

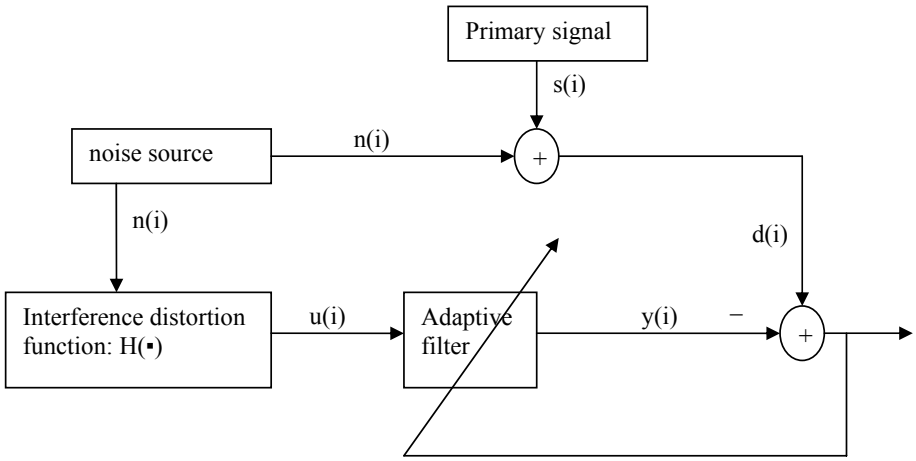

Fig. 3. The basic structure of the noise cancellation system

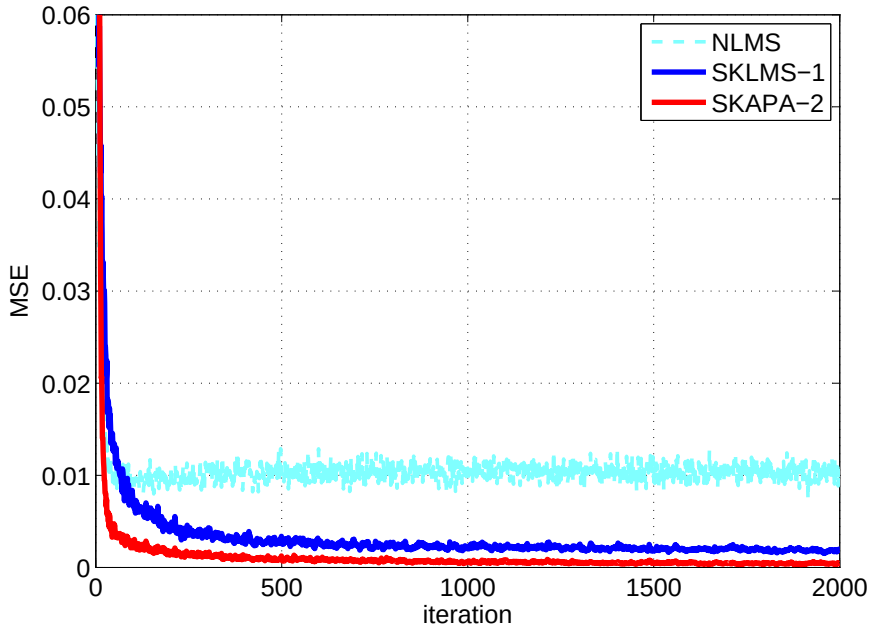

Fig. 4. Ensemble learning curves of NLMS, SKLMS-1 and SKAPA-2 ($K = 10$) in noise cancellation.

source $n(i)$. Therefore, the noise can be subtracted from $d(i)$ to improve the signal-noise-ratio.

In this example, the noise source is assumed white, uniformly distributed between $[-0.5, 05]$. The interference distortion function is assumed to be

$$u(i) = n(i) - 0.2u(i-1) - u(i-1)n(i-1) + 0.1n(i-1) + 0.4u(i-2) \tag{43}$$

As we see, the distortion function has infinite impulsive response, which on the other hand, means it is impossible

TABLE VII
NOISE REDUCTION COMPARISON IN NOISE CANCELLATION

Algorithm	Network Size	NR(dB)
NLMS	N/A	9.40
SKLMS-1	581	16.97
SKAPA-2	507	22.99

to recover $n(i)$ from a finite time delay embedding of $u(i)$. We rewrite the distortion function as

$$n(i) = u(i) + 0.2u(i-1) - 0.4u(i-2) + (u(i-1) - 0.1)n(i-1)$$

Therefore the present value of the noise source $n(i)$ not only depends on the reference noise measure $[u(i), u(i-1), u(i-2)]$, but it also depends on the previous value $n(i-1)$, which in turn depends on $[u(i-1), u(i-2), u(i-3)]$ and so on. It means we need a very long time embedding (infinite long theoretically) in order to recover $n(i)$ accurately. However, the recursive nature of the adaptive system provides a feasible alternative, i.e. we feedback the output of the filter $\hat{n}(i-1)$, which is the estimate of $n(i-1)$ to estimate the present one, pretending $\hat{n}(i-1)$ is the true value of $n(i-1)$. Therefore the input of the adaptive filter can be in the form of $[u(i), u(i-1), u(i-2), \hat{n}(i-1)]$. It can be seen that the system is inherently recurrent. In the linear case with a DARMA model, it is studied as *output error methods* [24]. However, it will be non-trivial to generalize the results concerning convergence and stability to nonlinear cases and we will address it in the future work.

We assume the primary signal $s(i) = 0$ during the training phase. And the system simply tries to reconstruct the noise source from the reference measure. We use a linear filter trained with the normalized LMS, two nonlinear filters trained with the SKLMS-1 and the SKAPA-2 ($K = 10$) respectively. 2000 training samples are used and 400 Monte C arlo simulations are run to get the ensemble learning curves as shown in Fig. 4. The step size and regularization parameter for the NLMS is 0.2 and 0.005. The step size for SKLMS-1 and SKAPA-2 is 0.5 and 0.2 respectively. The Gaussian kernel is used for both KLMS and KAPA with kernel parameter $a = 1$. The tolerance parameters for KLMS and KAPA are $\delta_1 = 0.15$ and $\delta_2 = 0.01$. And the noise reduction factor (NR), which is defined as $10\log_{10}\{E[n^2(i)]/E[n(i) - y(i)]^2\}$ is listed in Table VII. The performance improvement of SKAPA-2 is obvious when compared with SKLMS-1.

C. Nonlinear Channel Equalization

In this example, we consider a nonlinear channel equalization problem, where the nonlinear channel is modeled by a nonlinear Wiener model. The nonlinear Wiener model consists of a serial connection of a linear filter and a memoryless nonlinearity (See Fig. 5). This kind of model has been used to model digital satellite communication channels [25] and digital magnetic recording channels [26].

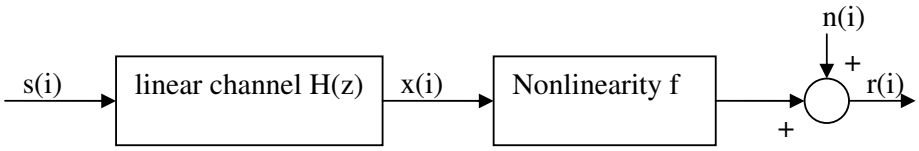

Fig. 5. Basic structure of the nonlinear channel.

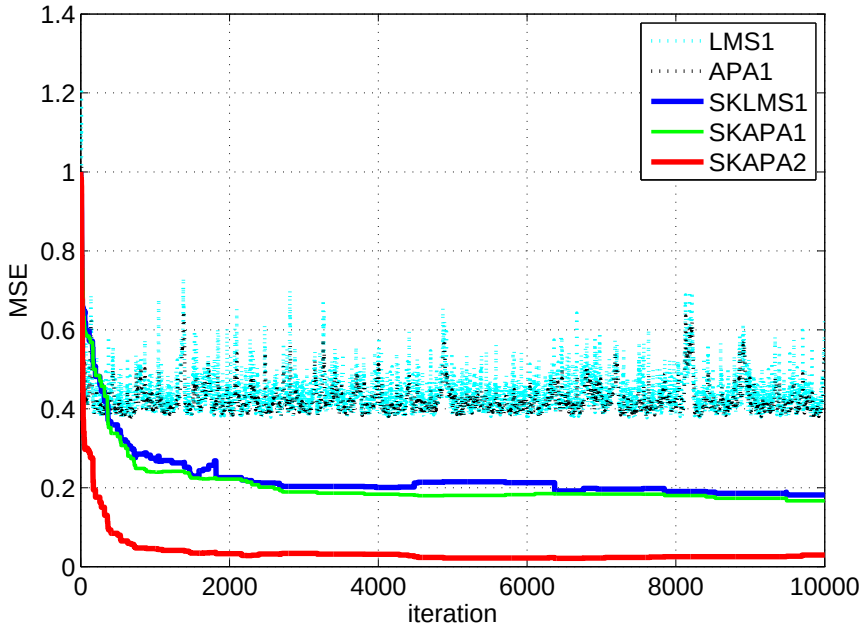

Fig. 6. The learning curves of the LMS1, APA1, SKLMS1, SKAPA1 and SKAPA2 in the nonlinear channel equalization ($\sigma = 0.1$).

The problem setting is as follows: A binary signal $\{s(1), s(2), ..., s(N)\}$ is fed into the nonlinear channel. At the receiver end of the channel, the signal is further corrupted by additive i.i.d. Gaussian noise and is then observed as $\{r(1), r(2), ..., r(N)\}$. The aim of channel equalization (CE) is to construct an *inverse* filter that reproduces the original signal with as low an error rate as possible. It is easy to formulate it as a regression problem, with input-output examples $\{(r(t+D), r(t+D-1), ..., r(t+D-l+1)), s(t)\}$, where l is the time embedding length, and D is the equalization time lag.

In this experiment, the nonlinear channel model is defined by $x(t) = s(t) + 0.5s(t-1)$, $r(t) = x(t) - 0.9x(t)^2 + n(t)$, where $n(t)$ is the white Gaussian noise with a variance of σ^2. We compare the performance of the LMS1, the APA1, the SKLMS1, the SKAPA1 ($K = 10$), and the SKAPA2 ($K = 10$). The Gaussian kernel with $a = 0.1$ is used in the SKLMS and SKAPA selected with cross validation. $l = 3$ and $D = 2$ in the equalizer. The noise variance is fixed here $\sigma = 0.1$. The learning curve is plotted in Fig. 6. The MSE is calculated between the continuous output (before taking the hard decision) and the desired signal. For the SKLMS1, SKAPA1, and SKAPA2, the novelty criterion is employed with $\delta_1 = 0.07$, $\delta_2 = 0.08$. The dynamic changing of the network size is also plotted in Fig. 7

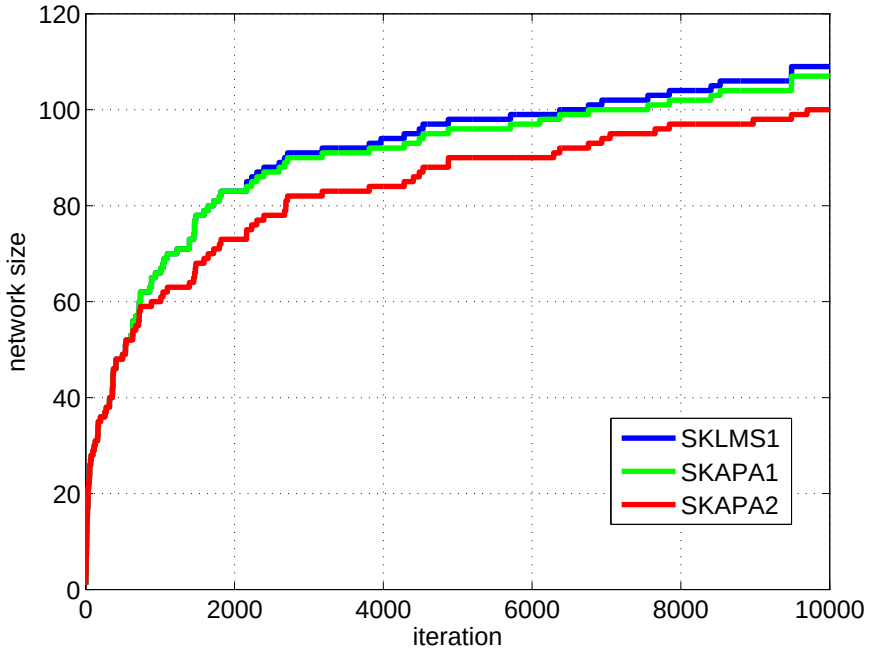

Fig. 7. Network size over training in the nonlinear channel equalization

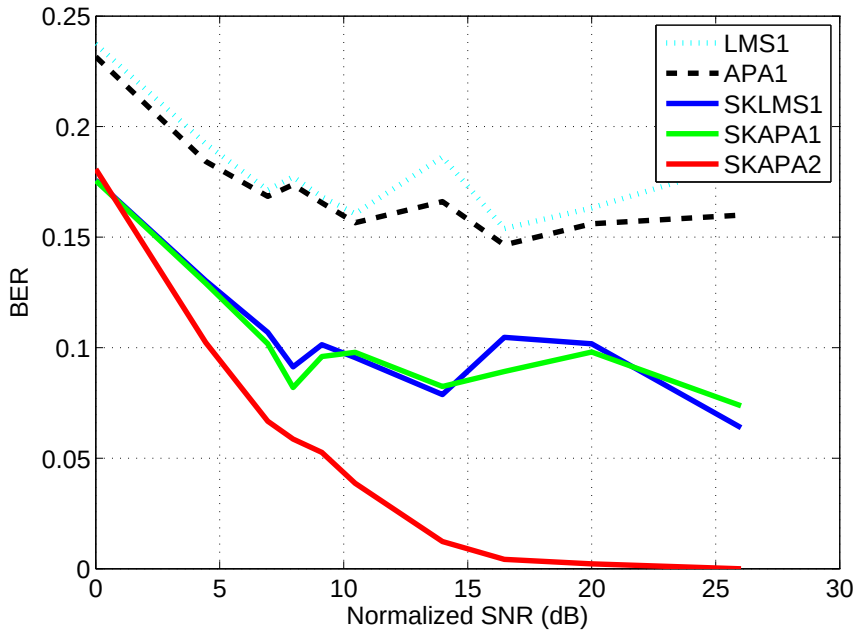

Fig. 8. Performance comparison with different SNR in the nonlinear channel equalization

over the training. It can be seen that at the beginning, the network sizes increase quickly but after convergence the network sizes increase slowly. And in fact, we can stop adding new centers after convergence by cross-validation by noticing that the MSE does not change after convergence.

Next, different noise variances are set. To make the comparison fair, we tune the novelty criterion parameters

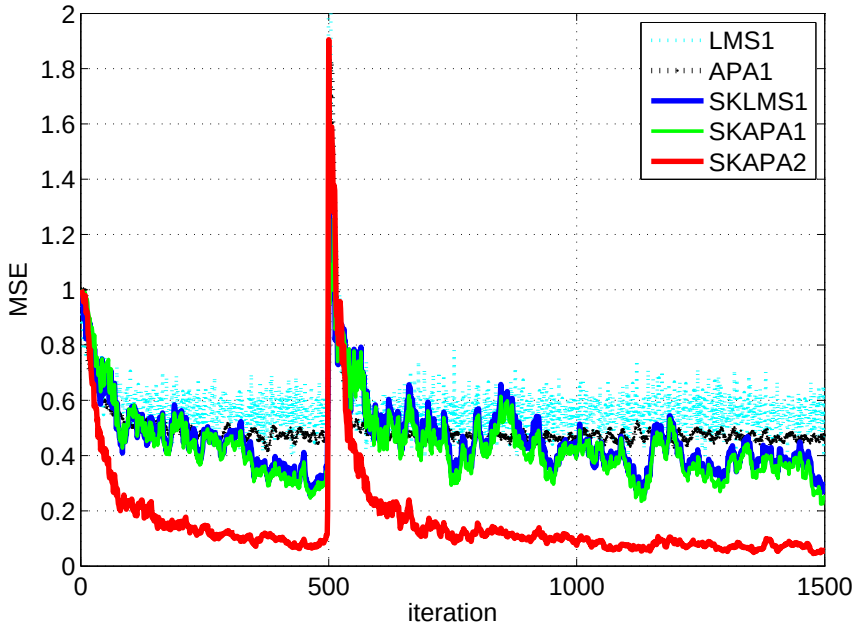

Fig. 9. Ensemble learning curves in the nonlinear channel equalization with an abrupt change at iteration 5000

to make the network size almost the same (around 100) in each scenario by cross validation. For each setting, 20 Monte Carlo simulations are run with different training data and different testing data. The size of the training data is 1000 and the size of the testing data is 10^5. The filters are fixed during the testing phase. The results are presented in Fig. 8. The normalized signal-noise-ratio (SNR) is defined as $10\log_{10}\frac{1}{\sigma^2}$. It is clearly shown that the SKAPA-2 outperforms the SKLMS-1 substantially in terms of the bit error rate (BER). The linear methods never really work in this simulation regardless of the SNR. The improvement of the SKAPA-1 on the SKLMS-1 is marginal but it exhibits a smaller variance. The roughness in the curves is mostly due to the variance from the stochastic training.

In the last simulation, we test the tracking ability of the proposed methods by introducing an abrupt change during training. The training data is 1500. For the first 500 data, the channel model is kept the same as before, but for the last 1000 data the nonlinearity of the channel is switched to $r(t) = -x(t) + 0.9x(t)^2 + n(t)$. The ensemble learning curves from 100 Monte Carlo simulations are plotted in Fig. 9 and the dynamic change of the network size is plotted in Fig. 10. It is seen that the SKAPA-2 outperforms other methods with its fast tracking speed. It is also noted that the network sizes increase right after the change to the channel model.

VII. Discussion and Conclusion

This paper proposes the KAPA algorithm family which is intrinsically a stochastic gradient methodology to solve the Least Squares problem in RKHS. It is a follow-up study of the recently introduced KLMS. Since the KAPA update equation can be written as inner products, KAPA can be efficiently computed in the input space. The good approximation ability of the KAPA stems from the fact that the transformed data $\varphi(\mathbf{u})$ includes possibly infinite

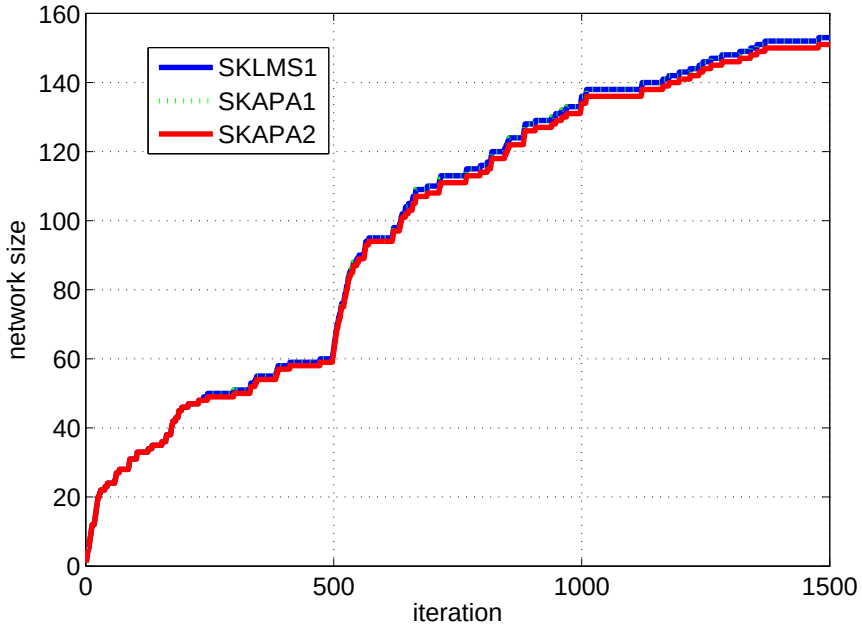

Fig. 10. Network size over training in the nonlinear channel equalization with an abrupt change at iteration 500

different features of the original data. In the framework of stochastic projection, the space spanned by $\varphi(\mathbf{u})$ is so large that the projection error of the desired signal could be very small [27], as is well known from Cover's theorem [28]. This capability includes modeling of nonlinear systems, which is the main reason why the KAPA can achieve good performance in the Mackey-Glass system prediction, adaptive noise cancellation, and nonlinear channel equalization.

Comparing with the KLMS, KRLS, and regularization networks (batch mode training), KAPA gives yet another way of calculating the coefficients for shallow RBF like neural networks. The performance of the KAPA is somewhere between the KLMS and KRLS, which is specified by the window length K. Therefore it not only provides a further theoretical understanding of RBF like neural networks, but it also brings much flexibility for application design with the constraints on performance and computation resources.

Three examples are studied in the paper, namely, time series prediction, nonlinear channel equalization and nonlinear noise cancellation. In all examples, the KAPA demonstrates superior performance when compared with the KLMS, which is expected from the classic adaptive filtering theory.

As pointed out, the study of the KLMS and KAPA has a close relation with the resource-allocating networks, but in the framework of RKHS, any Mercer kernel can be used instead of restricting the architecture to the Gaussian kernel. An important avenue for further research is how to choose the optimal kernel for a specific problem. A lot of work [29], [30], [31] has been done in the context of classical machine learning, which is usually derived in a strict optimization manner. Notice that with stochastic gradient methods, the solution obtained is not strictly the

optimal solution, therefore, further investigation is warranted. As we mentioned before, how to control the network size is still a big issue, which needs further study.

ACKNOWLEDGMENT

This work was partially supported by NSF grant ECS-0601271.

REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer Verlag, 1995.

[2] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[3] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, 2004.

[4] I. K. Kwang, M. Franz, and B. Scholkopf, "Iterative kernel principal component analysis for image modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1351–1366, September 2005.

[5] T.-T. Frieb and R. F. Harrison, "A kernel-based adaline," in *proceedings European Symposium on Artificial Neural Networks 1999*, April 1999, pp. 245–250.

[6] P. Pokharel, W. Liu, and J. Príncipe, "Kernel lms," in *Proc. International Conference on Accoustics, Speech and Signal Processing 2007*, 2007, pp. 1421–1424.

[7] J. Platt, "A resource-allocating network for function interpolation," *Neural Computation*, vol. 3, no. 2, pp. 213–225, 1991.

[8] A. Sayed, *Fundamentals of Adaptive Filtering*. New York: Wiley, 2003.

[9] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[10] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[11] B. Schölkopf, R. Herbrich, A. Smola, and R. Williamson, "A generalized representer theorem," in *Proceedings of the Annual Conference on Computational Learning Theory*, 2001, pp. 416–426.

[12] W. Liu, P. Pokharel, and J. C. Príncipe, "The kernel least mean square algorithm," *IEEE Transactions on Signal Processing*, vol. 56, 2008, to be published.

[13] F. Girosi, M. Jones, and T. Poggio, "Regularization theory and neural networks architectures," *Neural Computation*, vol. 7, pp. 219–269, 1995.

[14] S. V. Vaerenbergh, J. Via, and I. Santamaría, "A sliding-window kernel rls algorithm and its application to nonlinear channel identification," in *Proc. International Conference on Accoustics, Speech and Signal Processing 2006*, May 2006, pp. 789–792.

[15] J. Kivinen, A. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Trans. on Signal Processing*, vol. 52, pp. 2165–2176, Aug. 2004.

[16] W. Liu, P. Pokharel, and J. Príncipe, "Recursively adapted radial basis function networks and its relationship to resource allocating networks and online kernel learning," in *proceedings IEEE International workshop on machine learning for signal processing 2007*, 2007, pp. 245–250.

[17] G. Cavallanti, N. Cesa-Bianchi, and C. Gentile, "Tracking the best hyperplane with a simple budget perceptron," *Machine Learning*, vol. 69, pp. 143–167, 2007.

[18] Y. Sun, P. Saratchandran, and N. Sundararajan, "A direct link minimal resource allocation network for adaptive noise cancellation," *Neural processing letters*, vol. 12, no. 3, pp. 255–265, 2000.

[19] O. Dekel, S. Shalev-Shwartz, and Y. Singer, "The forgetron: A kernel-based perceptron on a fixed budget," in *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press, 2006, pp. 1342–1372.

[20] A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifiers with online and active learning," *J. Mach. Learn. Res.*, vol. 6, pp. 1579–1619, 2005.

[21] K. Fukumizu, "Active learning in multilayer perceptrons," in *Advances in Neural Information Processing Systems*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., vol. 8. The MIT Press, 1996, pp. 295–301.

[22] L. Glass and M. Mackey, *From Clocks to Chaos: The Rhythms of Life*. Princeton, NJ: Princeton University Press, 1988.

[23] S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines," in *IEEE Workshop on Neural Networks for Signal Processing VII*, J. Principe, L. Giles, N. Morgan, and E. Wilson, Eds. Piscataway, NJ: IEEE Press, 1997, pp. 511–514.

[24] G. C. Goodwin and K. S. Sin, *Adaptive filtering prediction and control*. Prentice Hall, 1984.

[25] G. Kechriotis, E. Zarvas, and E. S. Manolakos, "Using recurrent neural networks for adaptive communication channel equalization," *IEEE Trans. on Neural Networks*, vol. 5, pp. 267-278, March 1994.

[26] N. P. Sands and J. M. Cioffi, "Nonlinear channel models for digital magnetic recording," *IEEE Trans. Magn.*, vol. 29, pp. 3996-3998, November 1993.

[27] E. Parzen, "Statistical methods on time series by hilbert space methods," Applied Mathematics and Statistics Laboratory, Stanford University, CA, Tech. Rep. 23, 1959.

[28] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Prentice Hall, 1998.

[29] C. Micchelli and M. Pontil, "Learning the kernel function via regularization," *Journal of Machine Learning Research*, vol. 6, pp. 1099–1125, 2005.

[30] A. Argyriou, C. A. Micchelli, and M. Pontil, "Learning convex combinations of continuously parameterized basic kernels," in *Proc. of the 18th Conference on Learning Theory*, 2005, pp. 338–352.

[31] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, pp. 131–159, 2002.