Sparse Representation in Szegő Kernels through Reproducing Kernel Hilbert Space Theory with Applications

Y. Mo * T. Qian † W. Mi ‡

Abstract. This paper discusses generalization bounds for complex data learning which serve as a theoretical foundation for complex support vector machine (SVM). Drawn on the generalization bounds, a complex SVM approach based on the Szegő kernel of the Hardy space $H^2(\mathbb{D})$ is formulated. It is applied to the frequency-domain identification problem of discrete linear time-invariant system (LTIS). Experiments show that the proposed algorithm is effective in applications.

Keywords. sparse representation; support vector machine; Szegő kernel; reproducing kernel Hilbert space; frequency-domain identification.

1 INTRODUCTION

In learning from a set of examples, the key property of a learning algorithm is generalization, namely the ability of an algorithm to perform accurately on new examples after having trained on a set of training data. The training examples come from some generally unknown probability distribution, while the learner has to extract from them something more general that allows him to produce useful predictions in new cases. As for learning with real-valued examples for classification and regression, many researchers including Duda and Hart [10], John Shawe-Taylor and Nello Cristianini [33], Vapnik and Chervonenkis [40] devote themselves to working on the generalization of the learning algorithm. This paper takes into consideration the bounds on the generalization errors for complex-valued data. We obtain that the bounds on the generalized errors hold with high

*Department of Mathematics, University of Macau, Macao.*E-mail address*: marvel-2008@163.com.

† Corresponding author. Department of Mathematics, University of Macau, Macao. *E-mail address*: fsttq@umac.mo. The work was supported by Macao Sci. and Tech. Develop. Fund FDCT 098/2012/A3 and University of Macau research grant RC Ref No: MYRG116(Y1-L3)-FST13-QT.

‡ School of Mathematical Science, University of Electronic Science and Technology of China. *E-mail address*: mathmw@uestc.edu.cn.

probability over randomly chosen training set where the generalization errors are the probability of failing to achieve a target accuracy in complex-valued function prediction. This type of bounds allows one to conclude that the error is small with high probability. This result is an extension of John Shawe-Taylor and Nello Cristianini [33] to complex-valued data learning.

The generalization bounds give us a foundation to employ complex SVM. SVM is a system for efficiently training the linear learning machines in the kernel-induced feature spaces, with respect to the insights provided by the generalization theory. SVM was first proposed to obtain maximum margin separating hyperplanes in classification problems [39]. This technique now has become an integral part of a general learning theory [27, 32]. A comprehensive description of this method for classification and regression problems can be found in [4] and [34] respectively. It has been shown that SVMs perform well in practice in system identification, such as time series analysis with real SVM regression [23, 35] and channel estimation in LTE Downlink system with complex SVM regression based on the Gaussian kernel [5]. In this paper, we will construct a complex SVM with the Szegő kernel which is complex-valued and never be used in SVM framework before. This complex SVM will be applied to system identification.

System identification concerns the model of physical systems that can be described by input-output measurements in time domain or frequency domain. We concern the problem of approximating the dynamics of "single input, single output (SISO)" discrete-time linear time-invariant systems (LTIS) that are causal and stable in frequency domain. A discrete-time LTIS can be represented by [8, 24, 43]

$$f(z) = \sum_{k=0}^{\infty} h_k z^k, \tag{1}$$

where $f(z)$ is called the transfer function of the discrete-time LTIS, which is given by the Z-transform (generally evaluated at $1/z$). Researchers have been using the rational orthogonal systems by making the model structure priori-linear in parameters, viz, the transfer function $f(z)$ is approximated by

$$\sum_{k=1}^{n} \theta(k)\mathcal{B}_k(z),$$

where $\{\mathcal{B}_k\}$ is a rational orthogonal system, $\{\theta(k)\}$ is an $n-$tuple of parameters to be determined, n is the order of the model structure. A rational orthogonal system is also called a Takenaka-Malmquist (TM) system. In the unit disc $\mathbb{D}$ case, a TM system is defined by

$$\mathcal{B}_k(z) = \frac{\sqrt{1-|a_k|^2}}{1-\overline{a_k}z} \prod_{l=1}^{k-1} \frac{z-a_l}{1-\overline{a_l}z}, \tag{2}$$

where $a_k \in \mathbb{D}$ and $k = 1, 2, \ldots$.

For $k = 1$,

$$\mathcal{B}_1(z) = \frac{\sqrt{1-|a_1|^2}}{1-\overline{a_1}z} \tag{3}$$

is the normalized and parameterized (by a_1) Szegő kernel. We note that a TM system is generated from a sequence of parameterized Szegő and multi-Szegő kernels by using the Gram-Schmidt orthogonalization process [28].

It is known that [36] $\{\mathcal{B}_k\}_{k=1}^{\infty}$ forms a basis of Hardy space

$$H^2(\mathbb{D}) = \{f : f \text{ is analytic in } \mathbb{D}, \sup_{0<r<1} \int_0^{2\pi} |f(re^{i\theta})|^2 d\theta < \infty\},$$

if and only if

$$\sum_{k=1}^{\infty}(1-|a_k|) = \infty. \tag{4}$$

The condition (4) is assumed in the classical study of TM systems in system identification.

One usually uses TM systems with prescribed parameters as poles. Different choices of poles of a TM system give rise to different model structures, such as the classical FIR models (Fourier) by setting all poles being zero, the Laguerre models [18, 42] by setting all poles being equal to a fixed real number, the Kautz models [41] by setting all the poles being equal to a fixed complex number and the generalized model by fixing at least 3 different poles [1, 14, 25].

Classical identification methods present some limitations. Analysis of discrete-time measurements with atypical samples (outliers) is neither easy nor immediate, and it is usually achieved by heuristic or even visual inspection methods [16]. In general terms, classical methods require previous determination of model complexity or number of parameters in the model, and they are quite sensitive to wrong order choices [17].

Frequency-domain identification by using AFD (Adaptive Fourier decomposition) is a newly proposed model [22] that is based on [29, 37],. The novelty of AFD is to select the parameters a_ks for the TM system according to the measurements of the transfer function in a one-to-one manner based on the Maximal Selection Principle. Numerical examples in [22] show that the frequency-domain identification algorithm by using AFD has a better performance than the FIR and the Laguerre models. However, this method also needs previous determination of model complexity.

In this paper, we offer another direct use of the Szegő kernel for system identification. That is, the transfer function will be approximated by linear combinations of parameterized Szegő kernels. In view of the Gram-Schmidt orthogonalization process on partial fractions, our result is essentially again a TM system approximation. However, the latter is based on an approach different from the traditional ones in using TM systems. We will use complex SVM method for system identification.

Complex SVM, like real SVM, is based on statistical learning theory which seeks to minimize upper bounds of the generalization error consisting of the sum of the training

error and a confidence interval. This principle is different from the commonly used empirical risk minimization (ERM) principle which only minimizes the training error. Based on this, complex SVMs achieve better generalization performance than the learning algorithms employing the ERM principle. As consequence, complex SVMs can usually achieve better results. The use of robust cost functions in complex SVMs can decrease the effect of outliers. Training SVM is equivalent to solving a linearly constrained quadratic programming problem. In addition, in complex SVM the model complexity does not need to be previously fixed, but depends on the given measurements.

In Section 2, we will formulate a complex generalization bound, being appropriate for the complex SVM regression which will be proposed and developed in Section 3. Section 4 presents experiments with comparisons with important existing methods to test the proposed complex SVM. In Section 5, conclusions are drawn.

2 GENERALIZATION BOUND

A generalization result for real-valued functions is as follows [33]. We fix a target accuracy $\theta > 0$ and $0 < \gamma \leq \theta$. Consider a real-valued (hypothesis) function class $\mathcal{F}$ with domain X. For a function $g \in \mathcal{F}$ and a training point $(x_i, y_i) \in X \times \mathbb{R}$, we define

$$\xi((x_i, y_i), g, \theta, \gamma) = \xi_i = \max\{0, \mid g(x_i) - y_i \mid -(\theta - \gamma)\}.$$

This quantity is the amount by which $|g(x_i) - y_i|$ exceeds $\theta - \gamma$ on the point (x_i, y_i) or 0 if g is within $\theta - \gamma$ of the targeted value. This is, in fact, the ε insensitive loss function [9] given by

$$\mathcal{L}^{\varepsilon}(e_i) = \begin{cases} 0, & |e_i| \leq \varepsilon, \\ |e_i| - \varepsilon, & \varepsilon \leq |e_i|, \end{cases} \tag{5}$$

with $\varepsilon = \theta - \gamma$ and $e_i = g(x_i) - y_i$.

We can use other loss functions and apply to them the corresponding analysis. In this section, we only consider the case using ε insensitivity cost function (5).

For a training set $S = ((x_1, y_1), \ldots, (x_l, y_l))$, define the vector valued

$$\xi = \xi(S, g, \theta, \gamma) = (\xi_1, \ldots, \xi_l).$$

Note that $\xi_i > \gamma$ means the error of g on (x_i, y_i) is larger than θ.

Proposition 2.1 *[33] Let $\mathcal{F}$ be the set of real-valued linear functionals on a real-Hilbert space X that, in accordance with the Riesz representation theorem, is identical with the space X itself. Fix $\theta \in \mathbb{R}, \theta > 0$, and a probability distribution P on the space $X \times \mathbb{R}$. If we restrict the inputs to the ball $B(0, R) = \{x \in X \ : \ \|x\| \leq R\}$, then there is a constant c such that with probability at least $1 - \delta$ over randomly drawn training sets S of size l and for all γ, $0 < \gamma \leq \theta$, the probability that a function $g \in \mathcal{F}$ with its representation $\mathbf{w}$ in X has error larger than θ on a randomly chosen input is bounded by*

$$\epsilon(l, \delta, \gamma) = \frac{c}{l}\left(\frac{\| \mathbf{w} \|_2^2 \, R^2 + \| \xi \|_1^2 \log(1/\gamma)}{\gamma^2} \log^2 l + \log\frac{1}{\delta}\right). \tag{6}$$

In other words, with the notation

$$\mathrm{err}_P(g,\theta) = P\left(\{(x,y) \in X \times \mathbb{R} \ : \ |g(x)-y| \geq \theta\}\right),$$

there holds

$$P^l\left(\{S : \mathrm{err}_P(g,\theta)) \leq \epsilon(l,\delta,\gamma)\}\right) \geq 1-\delta, \tag{7}$$

where P^l is the product probability induced by P over $(X \times \mathbb{R})^l$.

There are other generalization bounds in terms of other norms of ξ [7, 33], but we restrict ourselves to the above norms in this paper. The above proposition is a theoretical foundation of the support vector regression for real-valued functions. Ignoring the logarithmic factor of the quantity to be minimized to improve the generalization, the support vector regression algorithm minimizes the quantity $\|\mathbf{w}\|_2^2 R^2 + \mathcal{C}\|\xi\|_1^2$ and hence optimizes the bound of Proposition 2.1.

We are able to prove a result of the same type for complex SVM regression algorithm.

Before we state our generalization bound theorem for complex Hilbert spaces and complex functionals, we need some preparations. Let X be a complex Hilbert space and g be a functional of X. The Reisz representation Theorem shows that g is induced by an element of X, say $\mathbf{w}$. That is $g(x) =< x, \mathbf{w} >$. Let $\{e_1, ..., e_n, ...\}$ be an arbitrary orthonormal basis of X. We define the "real" Hilbert space with respect to the basis as

$$x_R = \sum_{k=1}^{\infty} d_k e_k, \quad \sum_{k=1}^{\infty} {d_k}^2 < \infty, \quad d_k \in \mathbb{R}, \quad k = 1, 2, ...$$

For any $x \in X$ we have the decomposition

$$x = \sum_{k=1}^{\infty} c_k e_k = \sum_{k=1}^{\infty} \mathrm{Re}(c_k) e_k + i \sum_{k=1}^{\infty} \mathrm{Im}(c_k) e_k = x_R + i x_I, \ c_k \in \mathbb{C}.$$

We call the two infinite sums as the real and the complex parts of x, respectively. Similarly, let $\mathbf{w} = \mathbf{w}_R + i\mathbf{w}_I$. In such way, the role of $\mathbf{w}$ on X is split into four linear functionals on X_R :

$$g(x) = (\langle x_R, \mathbf{w}_R \rangle + \langle x_I, \mathbf{w}_I \rangle) + i(\langle x_R, -\mathbf{w}_I \rangle + \langle x_I, \mathbf{w}_R \rangle).$$

Let $g_R(x) = \langle x_R, \mathbf{w}_R \rangle + \langle x_I, \mathbf{w}_I \rangle$, $g_I(x) = \langle x_R, -\mathbf{w}_I \rangle + \langle x_I, \mathbf{w}_R \rangle$. Then $g(x) = g_R(x) + i g_I(x)$, $g_R(x)$ and $g_I(x)$ are linear and real-valued.
For $(x_i, y_i) \in X \times \mathbb{C}$, set

$$\xi((x_i, y_i), g, \theta/2, \gamma/2) = \xi_i = \max\{0, \mid g_R(x_i) - \mathrm{Re(y_i)}) \mid -(\theta-\gamma)/2\}$$

and

$$\xi^*((x_i, y_i), g, \theta/2, \gamma/2) = \xi_i^* = \max\{0, \mid g_I(x_i) - \mathrm{Im(y_i)} \mid -(\theta-\gamma)/2\}.$$

For a training set S, we define the corresponding real-vector valued

$$\xi = \xi(S, g, \theta/2, \gamma/2) = (\xi_1, \ldots, \xi_l), \quad \xi^* = \xi^*(S, g, \theta/2, \gamma/2) = (\xi_1^*, \ldots, \xi_l^*).$$

Note that $g_R(x)$ is a functional in the product space $X_R \times X_R$ of the tensor type, that is

$$g_R(x) = \langle (x_R, x_I), (\mathbf{w}_R, \mathbf{w}_I) \rangle = \langle x_R, \mathbf{w}_R \rangle + \langle x_I, \mathbf{w}_I \rangle,$$

similarly,

$$g_I(x) = \langle (x_R, x_I), (-\mathbf{w}_I, \mathbf{w}_R) \rangle = \langle x_R, -\mathbf{w}_I \rangle + \langle x_I, \mathbf{w}_R \rangle.$$

Then we have

Theorem 2.1 *Let $\mathcal{F}$ be the set of complex-valued linear functionals on a complex-Hilbert space X that, in accordance with the Riesz representation theorem, is identical with the space X itself. Fix $\theta \in \mathbb{R}, \theta > 0$, and a probability distribution P on the space $X \times \mathbb{C}$. If we restrict the inputs to the ball $B(0, R) = \{x \in X \ : \ \|x\| \leq R\}$, then there is a constant c such that with probability at least $1 - \delta$ over randomly drawn training sets S of size l and for all γ, $0 < \gamma \leq \theta$, the probability that a function $g \in \mathcal{F}$ with its representation $\mathbf{w}$ in X has error larger than θ on a randomly chosen input is bounded by*

$$\epsilon'(l, \delta, \gamma) = \frac{c}{l} \left(\frac{4(\| \mathbf{w} \|_2^2 \, R^2 + (\| \xi \|_1^2 + \| \xi^* \|_1^2) \log(2/\gamma))}{\gamma^2} \log^2 l + 2 \log \frac{2}{\delta} \right). \tag{8}$$

In other words, with the notation

$$\mathrm{err}_P(g, \theta) = P\left(\{(x, y) \in X \times \mathbb{C} \ : \ |g(x) - y| \geq \theta\} \right),$$

there holds

$$P^l \left(\{S : \mathrm{err}_P(g, \theta)) \leq \epsilon'(m, \delta, \gamma)\} \right) \geq 1 - \delta, \tag{9}$$

where P^l is the product probability induced by P over $(X \times \mathbb{C})^l$.

Proof. We observe that in the space $X \times \mathbb{C}$, $\forall g \in \mathcal{F}$, we have

$$\{|g(x) - y| > \theta\} \subset \{|g_R(x) - \mathrm{Re(y)}| > \theta/2\} \subset \{|g_{\mathrm{I}}(\mathrm{x}) - \mathrm{Im(y)}| > \theta/2\}.$$

Denote by $P_{(X_R \times X_R) \times \mathbb{R}}$ the induced probability on $(X_R \times X_R) \times \mathbb{R}$ from P on $X \times \mathbb{C}$ and

$$\mathrm{err}_{P_{(X_R \times X_R) \times \mathbb{R}}}(g_R, \theta, \xi) = P_{(X_R \times X_R) \times \mathbb{R}}(\{(x, \mathrm{Re(y)}) \in (\mathrm{X_R} \times \mathrm{X_R}) \times \mathbb{R} \ : \ |g_{\mathrm{R}}(\mathrm{x}) - \mathrm{Re(y)}| > \theta\}),$$

where ξ on the left-hand-side is used as slack variables. By Proposition 2.1, for some $c = c_R$ in the relation (6) there holds, for $|S| = l$,

$$P^l_{(X_R \times X_R) \times \mathbb{R}}(\{S : \mathrm{err}_{P_{(X_R \times X_R) \times \mathbb{R}}}(g_R, \theta/2, \xi) > \epsilon_{g_R, \xi, c_R}(l, \delta/2, \gamma/2)\}) < \delta/2,$$

where the ϵ_{g_R, ξ, c_R} is defined similarly in (6) but here with the dependence of g_R, ξ and c_R. Similarly, there holds

$$P^l_{(X_R \times X_R) \times \mathbb{R}}(\{S : \mathrm{err}_{P_{(X_R \times X_R) \times \mathbb{R}}}(g_I, \theta/2, \xi^*) > \epsilon_{g_I, \xi^*, c_I}(l, \delta/2, \gamma/2)\}) < \delta/2,$$

Let $c = \max\{c_R, c_I\}$. Then we have

$$P^l_{(X_R\times X_R)\times\mathbb{R}}(\{S : \mathrm{err}_{P_{(X_R\times X_R)\times\mathbb{R}}}(g_R, \theta/2, \xi) > \epsilon_{g_R,\xi,c}(l, \delta/2, \gamma/2)\}) < \delta/2,$$

and

$$P^l_{(X_R\times X_R)\times\mathbb{R}}(\{S : \mathrm{err}_{P_{(X_R\times X_R)\times\mathbb{R}}}(g_I, \theta/2, \xi^*) > \epsilon_{g_I,\xi^*,c}(l, \delta/2, \gamma/2)\}) < \delta/2.$$

Define

$$\epsilon'(l, \delta, \gamma) = \epsilon_{g_R,\xi,c}(l, \delta/2, \gamma/2) + \epsilon_{g_I,\xi^*,c}(l, \delta/2, \gamma/2).$$

Since

$$\mathrm{err}_P(g, \theta) \le \mathrm{err}_{P_{(X_R\times X_R)\times\mathbb{R}}}(g_R, \theta/2, \xi) + \mathrm{err}_{P_{(X_R\times X_R)\times\mathbb{R}}}(g_I, \theta/2, \xi^*),$$

we have that the set

$$\{(x, y) \in X \times \mathbb{C} \ : \ \mathrm{err}_P(g, \theta) > \epsilon'(l, \delta, \gamma)\}$$

is contained in the following

$$\begin{aligned}&\{(x, \mathrm{Re(y)}) \in (\mathrm{X_R} \times \mathrm{X_R}) \times \mathbb{R} \ : \ \mathrm{err}_{\mathrm{P}_{(\mathrm{X_R}\times \mathrm{X_R})\times\mathbb{R}}}(g_\mathrm{R}, \theta/2, \xi) > \epsilon_{g_\mathrm{R},\xi,\mathrm{c}}(l, \delta/2, \gamma/2)\}\\ &\cup\{(x, \mathrm{Im(y)}) \in (\mathrm{X_R} \times \mathrm{X_R}) \times \mathbb{R} \ : \ \mathrm{err}_{\mathrm{P}_{(\mathrm{X_R}\times \mathrm{X_R})\times\mathbb{R}}}(g_\mathrm{I}, \theta/2, \xi^*) > \epsilon_{g_\mathrm{I},\xi^*,\mathrm{c}}(l, \delta/2, \gamma/2)\}.\end{aligned}$$

Therefore, on randomly chosen training sets S of cardiality l, for a ramdomly chosen input, we have, in the product space,

$$\begin{aligned}&P^l_{X\times\mathbb{C}}(\{S : \ \mathrm{err}_P(g, \theta) > \epsilon'(l, \delta, \gamma)\}) \le\\ &P^l_{(X_R\times X_R)\times\mathbb{R}}(\{S : \mathrm{err}_{P_{X_R\times X_R)\times\mathbb{R}}}(g_R, \theta/2, \xi) > \epsilon_{g_R,\xi,c}(l, \delta/2, \gamma/2)\}) +\\ &P^l_{(X_R\times X_R)\times\mathbb{R}}(\{S : \mathrm{err}_{P_{X_R\times X_R)\times\mathbb{R}}}(g_I, \theta/2, \xi^*) > \epsilon_{g_I,\xi^*,c}(l, \delta/2, \gamma/2)\}) < \delta.\end{aligned}$$

The desired result then follows. □

Remark 2.1 *Ignoring the logarithmic factor of the quantity to be minimized to improve the generalization, the complex support vector regression algorithm minimizes the objective function* $\|\mathbf{w}\|_2^2 R^2 + \mathcal{C}(\|\xi\|_1^2 + \|\xi^*\|_1^2)$ *in the input space* X *and hence optimizes the bound of Theorem 2.1. This Theorem is a theoretical foundation for linear complex-valued support vector machine. In the nonlinear setting, the input space should be the feature space and minimizes the functional* $\|\mathbf{w}\|_2^2 R^2 + \mathcal{C}(\|\xi\|_1^2 + \|\xi^*\|_1^2)$ *in feature space.*

3 SUPPORT VECTOR MACHINE FOR LTIS

Complex support vector machine can be used to approximate function by linear combinations of the parameterized reproducing kernels in complex reproducing kernel Hilbert spaces. In this section, we will construct a complex SVM for frequency-domain identification for LTIS.

3.1 PROBLEM SETTING

Frequency-domain identification is based on a set of frequency-domain measurements. Without loss of generality, it is assumed that a set of frequency-domain measurements $\{E_k\}_{k=1}^M$ are obtained from a single input, single output (SISO), discrete-time LTIS $f(z)$ in $H^2(\mathbb{D})$. It is further assumed that $f(z)$ can be continuously extended to a region containing the closed unit disc.

We introduce some notations here. Assume that the measurements $\{E_k\}_{k=1}^N$ are set to be

$$E_k = f(e^{jw_k}) \ (k = 1, 2, \ldots, N) \tag{10}$$

for the noiseless case; and,

$$E_k = f(e^{jw_k}) + v_k \ (k = 1, 2, \ldots, N) \tag{11}$$

for the noised case, where $w_k = \frac{2\pi(k-1)}{N-1}$ and $\{v_k\}$ can be either a bounded sequence with $|v_k| \leq \epsilon$, $\epsilon > 0$; or a zero mean stochastic process with a bounded covariance function. The noise we deal with is considered to be of the same property. Since h_ks in (1) are real-valued, we have $f(e^{jt}) = \overline{f(e^{j(2\pi - t)})}$. Therefore, in practice one just needs to measure the data for $w_k \in [0, \pi)$, and the rest in the interval $[\pi, 2\pi)$ can be set to be their conjugates.

The followings are some properties in relation to the reproducing kernel Hilbert space (RKHS) [2].

A RKHS is a Hilbert space of functions on some set X such that all evaluation functionals, i.e., the maps $f \longmapsto f(x)(x \in X)$, are continuous. In that case, by the Riesz representation theorem, for each $x \in X$, there exists a unique function, called K_x, such that

$$f(x) = \langle f, K_x \rangle, \tag{12}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of the Hilbert space. We also write $K_x(\cdot)$ as $K(\cdot, x)$, called parameterized reproducing kernel. $K(\cdot, \cdot)$ itself is the reproducing kernel of the Hilbert space.

Note that $\langle f, K_x(\cdot) \rangle = 0$ for all x implies that f is identically zero. Hence the set of functions $\{K_x(\cdot) : x \in X\}$ spans the whole RKHS. The dot product on the RKHS thus only needs to be defined on $\{K_x(\cdot) : x \in X\}$ and can then be extended to the whole RKHS by linearity and continuity.

From (12), it follows that in particular,

$$\langle K_y(\cdot), K_x(\cdot) \rangle = K(x, y) \tag{13}$$

for all $x, y \in X$.

Reproducing kernel can be referred to as the Mercer kernel in the SVM community [31]. The complex Hardy Space on the unit disc, $H^2(\mathbb{D})$, is, as well known, a RKHS with the reproducing kernel, Szegő kernel [12],

$$K(y, x) = \frac{1}{1 - \overline{x}y}, \tag{14}$$

where $x, y \in \mathbb{D}$.

The identification problem that we consider now can be stated as follows.

Frequency-domain identification problem: Given a set of noiseless or noise corrupted frequency-domain measurements $\{E_k\}_{k=1}^N$ for $f \in H^2(\mathbb{D})$. Find an M-linear combination of parameterized Szegő kernels, $f_M(z) = \sum_{k\in J} \psi_k K(z, z_k) \in H^2(\mathbb{D}), z, z_k \in \mathbb{D}$ and $|J| = M$ $(M \leq N)$ to reconstruct the system approximately.

3.2 ALGORITHM

There are three steps in the proposed algorithm. First, we work out a function $\tilde{f}(z) \in H^2(\mathbb{D})$ approximating the true function $f(z)$ depending on the given measurements $\{E_k\}_{k=1}^N$. Then we choose n $(n \leq N)$ samples $\{(z_m, y_m)\}_{m=1}^n$ $(z_m \in \mathbb{D})$ satisfying $y_m = \tilde{f}(z_m)$ $(m = 1, \ldots, n.)$. Based on the n samples chosen in the second step, we construct a complex SVM to get an approximating function to $\tilde{f}(z)$ in the form of a finite linear combination of parameterized Szegő kernels, as the third step.

3.2.1 CONSTRUCT $\tilde{f}(z)$

Assume that the system function f is complex analytic inside the unit disc and of finite energy restricted to the unit circle. The first step is to construct by using data $\{E_k\}_{k=1}^N$, the first approximation $\tilde{f}$, which is also analytic in the disc. By the same method as in [22], $\tilde{f}(z)$ is constructed by the Cauchy integral

$$\tilde{f}(z) = \frac{1}{2\pi j} \int_0^{2\pi} \frac{\sum_k E_k \chi_k(w)}{e^{jw} - z} de^{jw}, \tag{15}$$

where $\chi_k = \chi_{(w_k, w_{k+1})}$ is the indicator function. The above is identical with

$$\frac{1}{2\pi} \sum_{k=1}^N [f(e^{jw_k}) + v_k] \int_{w_k}^{w_{k+1}} \frac{e^{jw} dw}{e^{jw} - e^{jw_k}} \approx \sum_{k=1}^N [f(e^{jw_k}) + v_k] \frac{(w_{k+1} - w_k)}{2\pi} e^{jw_k},$$

being a discrete approximation of the Cauchy integral of $f(e^{iw})$ with the error, where $E_k = f(e^{jw_k}) + v_k$ stands for the measurements with or without noise (in the latter case $v_k = 0$). It is easy to show that $\tilde{f}(\bar{z}) = \overline{\tilde{f}(z)}$.

Remark 3.1 *To get an approximating function $\tilde{f}(z)$, use of the (15) can be avoided. Other approximation methods can also work for this purpose, such as by polynomials according to [13, 26].*

Usually, the difference between f and $\tilde{f}$ collects both the noise and the approximation errors. The role of $\tilde{f}$ is to get the approximation values at the training points.

Remark 3.2 *It is noted that the inner product in the Hardy space $H^2(\mathbb{D})$ is given by an integral over the boundary of the unit disc, viz., the unit circle:*

$$\langle f, g\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(e^{it})\overline{g(e^{it})}dt.$$

The Hardy space functions, however, are defined inside the disc. The above representation of inner product is based on the fact that a Hardy space function uniquely and isometrically corresponds to an L^2 -function: Its non-tangential limit on the boundary. Precisely speaking, each $f \in H^2(\mathbb{D})$ has a non-tangential boundary limit function $\hat{f} \in L^2(\partial\mathbb{D})$ such that $\|f\|_{H^2} = \|\hat{f}\|_{L^2}$ (see, for instance, [12]); on the other hand, the Cauchy integral of $\hat{f}, C(\hat{f})$, being a Hardy space function, is identical with f. The last assertion is proved by invoking the fact $H(\hat{f}) = -i\hat{f}$ for boundary limits of the Hardy space function, where $H(\hat{f})$ is the Hilbert transform of $\hat{f}$, and the Plemelj theorem

$$\lim_{z\to e^{it}\ nontangentially} C(\hat{f})(z) = \frac{1}{2}\hat{f}(e^{it}) + i\frac{1}{2}H(\hat{f})(e^{it}) = \hat{f}(e^{it}), \qquad a.e.,$$

as well as the fact that if two Hardy space functions have a.e. identical boundary limits, then they have to be identical.

Based on the above relations, the definition of the reproducing kernel of the Hardy space, $K(q,p) = \frac{1}{1-\bar{p}q}$, may be extended to $\mathbb{D}\times\bar{\mathbb{D}}$, and even to $\bar{\mathbb{D}}\times\bar{\mathbb{D}}$ in the sense that for $q, \tilde{q} \in \bar{\mathbb{D}}$,

$$\langle f, K_{\tilde{q}}\rangle = \lim_{q\to\tilde{q}\ nontangentially}\langle f, K_q\rangle = \lim_{q\to\tilde{q}\ nontangentially} f(q) = f(\tilde{q}), \quad a.e.$$

Based on the above noted, in our SVM approach, we can use variables in both the unit disc and on its boundary.

Remark 3.3 *Although the algorithm assumes that the data is from a function in the Hardy space, in practice, however, it will not be the case. Most data would be from functions in the L^2 space viz., from functions of finite energy. The theory and algorithm offered by this paper are valid to data from L^2 spaces as well. In fact, whenever the inner product is used, or a discretization of the inner product calculation is involved, we have, due to the orthogonality property between the two opposite Hardy spaces,*

$$\langle f, K_q\rangle = \langle f^+, K_q\rangle + \langle f^-, K_q\rangle = \langle f^+, K_q\rangle,$$

where $f = f^+ + f^-$, f^+ and f^- belong to, respectively, the so called Hardy spaces H^+ and H^-. If K_q is replaced by any other function in the Hardy space inside the disc, there holds a similar result. As consequence, what we obtain in the end of the program is an approximation to the Hardy space projection f^+ of f, not to f itself. In the case a real-valued function f of finite energy may be recovered by the relation

$$f = 2Ref^+ - c_0,$$

where c_0 is the average of the function on the circle, or the 0-th Fourier coefficient of f.

3.2.2 CHOOSE n POINTS FROM $\tilde{f}(z)$ as training points

Choose n ($n \leq N$) samples $\{(z_m, y_m)\}_{m=1}^n$ ($z_m \in \mathbb{D}$) satisfying $y_m = \tilde{f}(z_m)$ ($m = 1, \ldots, n$).

Remark 3.4 *Selection of samples in Step 3.2.2 is with great flexibility. The n samples can be taken evenly in the disc, which is commonly thought about. This thought is correct and supported by experiments in Section 4. On the other hand, according to the uniqueness theorem of analytic functions, if the function values of two analytic functions coincide at $z_1, z_2, \ldots, z_n, \ldots$, tending to a limit in the intersection of their analytic domains, then the two functions have to be identical. Therefore, distributions of sampling points do not have to be even in the disc.*

3.2.3 SVM FOR LTIS

In this section, based on the n samples chosen in Section 3.2.2, we construct a complex SVM to obtain an approximating function $f_M(z)$. In the following text, unless otherwise stated, K stands for the Szegő kernel.

Given $\{(z_m, y_m), m = 1, \ldots, n\}$, the problem consists of finding an approximating function g that fits the data as follows:

$$y_m = \tilde{f}(z_m) = g(z_m) + e_m \quad (m = 1, \ldots, n), \tag{16}$$

where residuals $\{e_m\}$ account for the approximation errors.

We start with a conventional SVM nonlinear regression [39]. We first map a given point set $\{z_m\}$ to the set $\{K_{z_m}(\cdot)\}$ in the higher dimensional space $H^2(\mathbb{D})$, i.e., the feature space, by using the transformation $\Phi : \mathbb{D} \to H^2 : z_m \to \Phi(z_m) = K_{z_m}(\cdot) = K(\cdot, z_m) \in H^2(\mathbb{D})$. In $H^2(\mathbb{D})$, a linear bounded functional itself is given by a function in $H^2(\mathbb{D})$, that is,

$$y_m = \tilde{f}(z_m) = g(z_m) + e_m = \langle \Phi(z_m), \mathbf{w} \rangle + e_m \quad (m = 1, \ldots, n), \tag{17}$$

where $\mathbf{w} \in H^2(\mathbb{D})$ is to be determined.

In accordance with a variation of Theorem 2.1, to get an optimal generalization bound is to minimize $\|\mathbf{w}\|^2 + \mathcal{C} \sum_{m=1}^{n} \mathcal{L}^{\varepsilon}(e_m)$, where we define $\mathcal{L}^{\varepsilon}(e_m) = \mathcal{L}^{\varepsilon}\mathcal{R}(e_m) + \mathcal{L}^{\varepsilon}\mathcal{I}(e_m)$ for complex e_m, $\mathcal{L}^{\varepsilon}(\cdot)$ is the ε insensitive cost function, $\mathcal{R}(\cdot)$ and $\mathcal{I}(\cdot)$ represent real and imaginary parts, respectively. ε is replacement of $(\theta - \gamma)/2$ in Theorem 2.1.

In this formulation, we adopt a more general cost function [20,30], which additionally considers a quadratic cost zone, called ε-Huber cost function, given by

$$\mathcal{L}^{\varepsilon}(e_m) = \begin{cases} 0, & |e_m| \leq \varepsilon, \\ \frac{1}{2r}(|e_m| - \varepsilon)^2, & \varepsilon \leq |e_m| \leq e_C, \\ C(|e_m| - \varepsilon) - \frac{1}{2}rC^2, & e_C \leq |e_m|, \end{cases} \tag{18}$$

where $e_C = \varepsilon + rC$, ε is an error-tolerance parameter for the training samples, r and C are free parameters that control the shape of the cost function.

By suitable choices of free parameters ε, r, C, the ε-Huber cost function can be adapted to different kinds of noise. The choice of $\mathcal{C}$ is absorbed in the choice of parameters ε, r, C. The primal problem is

$$\min \quad \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{m=1}^{n} \mathcal{L}^{\varepsilon}(e_m). \tag{19}$$

Define

$$\begin{aligned}
\xi_m &= \max\{0, \mathcal{R}(e_m) - \varepsilon\}, \\
\xi_m^* &= \max\{0, \mathcal{R}(-e_m) - \varepsilon\}, \\
\zeta_m &= \max\{0, \mathcal{I}(e_m) - \varepsilon\}, \\
\zeta_m^* &= \max\{0, \mathcal{I}(-e_m) - \varepsilon\},
\end{aligned}$$

Then we have

$$\mathcal{L}^{\varepsilon}\mathcal{R}(e_m) \quad = \mathcal{L}^{\varepsilon}\mathcal{R}(e_m)\chi_+(\mathcal{R}(e_m)) + \mathcal{L}^{\varepsilon}\mathcal{R}(e_m)\chi_-(\mathcal{R}(e_m)), \tag{20}$$

where $\chi_+ = \chi_{(0,+\infty)}$ and $\chi_- = \chi_{(-\infty,0)}$, for $m = 1, \ldots, n$.

If $\mathcal{R}(e_m) > 0$, the cost function can be written in the new notation as

$$\mathcal{L}^{\varepsilon}\mathcal{R}(e_m) = \begin{cases} 0, & \xi_m = 0, \\ \frac{1}{2r}\xi_m^2, & 0 \le \xi_m \le rC, \\ C\xi_m - \frac{1}{2}rC^2, & rC \le \xi_m. \end{cases} \tag{21}$$

Therefore,

$$\mathcal{L}^{\varepsilon}\mathcal{R}(e_m)\chi_+(\mathcal{R}(e_m)) = \frac{1}{2r}\xi_m^2\chi_{[0,rC]}(\xi_m) + (C\xi_m - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m).$$

Similarly,

$$\mathcal{L}^{\varepsilon}\mathcal{R}(e_m)\chi_-(\mathcal{R}(e_m)) = \frac{1}{2r}\xi_m^{*2}\chi_{[0,rC]}(\xi_m^*) + (C\xi_m^* - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m^*).$$

There is a similar representation for $\mathcal{L}^{\varepsilon}\mathcal{I}(e_m)$. Therefore, we have

$$\begin{aligned}
\mathcal{L}^{\varepsilon}(e_m) =& \mathcal{L}^{\varepsilon}\mathcal{R}(e_m) + \mathcal{L}^{\varepsilon}\mathcal{I}(e_m) \\
=& \frac{1}{2r}\xi_m^2\chi_{[0,rC]}(\xi_m) + (C\xi_m - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m) \\
&+ \frac{1}{2r}\xi_m^{*2}\chi_{[0,rC]}(\xi_m^*) + (C\xi_m^* - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m^*) \\
&+ \frac{1}{2r}\zeta_m^2\chi_{[0,rC]}(\zeta_m) + (C\zeta_m - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\zeta_m) \\
&+ \frac{1}{2r}\zeta_m^{*2}\chi_{[0,rC]}(\zeta_m^*) + (C\zeta_m^* - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\zeta_m^*).
\end{aligned}$$

Then we can give the primal problem (19) as (this is an extension of linear OFDM-SVM [11] to nonlinear LTIS SVM scenarios)

$$\begin{aligned}\min\ & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2r}\sum_{m=1}^{n}\xi_m^2\chi_{[0,rC]}(\xi_m) + \sum_{m=1}^{n}(C\xi_m - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m) \\ & + \frac{1}{2r}\sum_{m=1}^{n}\xi_m^{*2}\chi_{[0,rC]}(\xi_m^*) + \sum_{m=1}^{n}(C\xi_m^* - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m^*) \\ & + \frac{1}{2r}\sum_{m=1}^{n}\zeta_m^2\chi_{[0,rC]}(\zeta_m) + \sum_{m=1}^{n}(C\zeta_m - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\zeta_m) \\ & + \frac{1}{2r}\sum_{m=1}^{n}\zeta_m^{*2}\chi_{[0,rC]}(\zeta_m^*) + \sum_{m=1}^{n}(C\zeta_m^* - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\zeta_m^*)\end{aligned} \tag{22}$$

constrained to

$$\begin{aligned}\mathcal{R}(y_m - \langle\Phi(z),\mathbf{w}\rangle) &\le \varepsilon + \xi_m, \\ \mathcal{R}(-y_m + \langle\Phi(z),\mathbf{w}\rangle) &\le \varepsilon + \xi_m^*, \\ \mathcal{I}(y_m - \langle\Phi(z),\mathbf{w}\rangle) &\le \varepsilon + \zeta_m, \\ \mathcal{I}(-y_m + \langle\Phi(z),\mathbf{w}\rangle) &\le \varepsilon + \zeta_m^*, \\ \xi_m, \xi_m^*, \zeta_m, \zeta_m^* &\ge 0,\end{aligned} \tag{23}$$

for $m = 1, \ldots, n$.

The key idea is to construct a Lagrangian function from the primal functional and the corresponding constraints by introducing a Lagrangian multiplier (or dual variable) for each constraint of the primal problem. It can be shown that this function has a saddle point with respect to the primal and dual variables at the solution. For details, see e.g. [19, 21]. By including linear constraints (23) in (22), the Lagrangian function is obtained:

$$
\begin{aligned}
L =& \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2r}\sum_{m=1}^{n}\xi_m^2\chi_{[0,rC]}(\xi_m) + \sum_{m=1}^{n}(C\xi_m - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m) \\
&+\frac{1}{2r}\sum_{m=1}^{n}\xi_m^{*2}\chi_{[0,rC]}(\xi_m^*) + \sum_{m=1}^{n}(C\xi_m^* - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\xi_m^*) \\
&+\frac{1}{2r}\sum_{m=1}^{n}\zeta_m^2\chi_{[0,rC]}(\zeta_m) + \sum_{m=1}^{n}(C\zeta_m - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\zeta_m) \\
&+\frac{1}{2r}\sum_{m=1}^{n}\zeta_m^{*2}\chi_{[0,rC]}(\zeta_m^*) + \sum_{m=1}^{n}(C\zeta_m^* - \frac{1}{2}rC^2)\chi_{(rC,+\infty)}(\zeta_m^*) \\
&-\sum_{m=1}^{n}(\lambda_m\xi_m + \lambda_m^*\xi_m^*) - \sum_{m=1}^{n}(\eta_m\zeta_m + \eta_m^*\zeta_m^*) \\
&+\sum_{m=1}^{n}\alpha_m[\mathcal{R}(y_m - \langle\Phi(z),\mathbf{w}\rangle) - \varepsilon - \xi_m] + \sum_{m=1}^{n}\alpha_m^*[\mathcal{R}(-y_m + \langle\Phi(z),\mathbf{w}\rangle) - \varepsilon - \xi_m^*] \\
&+\sum_{m=1}^{n}\beta_m[\mathcal{I}(y_m - \langle\Phi(z),\mathbf{w}\rangle) - \varepsilon - \zeta_m] + \sum_{m=1}^{n}\beta_m^*[\mathcal{I}(-y_m + \langle\Phi(z),\mathbf{w}\rangle) - \varepsilon - \zeta_m^*].
\end{aligned} \tag{24}
$$

The lagrangian multipliers are constrained to

$$
\alpha_m^{(*)}, \beta_m^{(*)}, \lambda_m^{(*)}, \eta_m^{(*)} \geq 0 \quad (\alpha^{(*)} \text{ stand for } \alpha \text{ and } \alpha^*). \tag{25}
$$

We should also notice

$$
\xi_m^{(*)}, \zeta_m^{(*)} \geq 0. \tag{26}
$$

The following additional constraints must also be fulfilled:

$$
\alpha_m\alpha_m^* = 0, \beta_m\beta_m^* = 0. \tag{27}
$$

Besides this, the Karush-Kuhn-Tucker (KKT) conditions [39] yield

$$
\lambda_m\xi_m = 0,\ \lambda_m^*\xi_m^* = 0,\ \text{and}\ \ \eta_m\zeta_m = 0,\ \eta_m^*\zeta_m^* = 0. \tag{28}
$$

We have to minimize functional (24) with respect to the primal variables and then maximize it with respect to the dual variables under the constraints (25), (26), (27) and (28). To take the partial derivative, we employ the rules of Wirtinger′s Calculus for the complex variables on complex RKHS's as described in [3]. We have

$$
\frac{\partial L}{\partial \mathbf{w}^*} = \frac{1}{2}\mathbf{w} - \frac{1}{2}\sum_{m=1}^{n}\alpha_m\Phi(z_m) + \frac{1}{2}\sum_{m=1}^{n}\alpha_m^*\Phi(z_m) + \frac{j}{2}\sum_{m=1}^{n}\beta_m\Phi(z_m) - \frac{j}{2}\sum_{m=1}^{n}\beta_m^*\Phi(z_m).
$$

If the gradient is set to be zero, we obtain

$$\mathbf{w} = \sum_{m=1}^{n} \psi_m \Phi(z_m) = \sum_{m=1}^{n} \psi_m K_{z_m}(\cdot), \tag{29}$$

where $\psi_m = (\alpha_m - \alpha_m^*) - j(\beta_m - \beta_m^*)$.

For the real variables, we compute the gradients in the traditional way. If the gradients of L_{pd} with respect to $\xi_m^{(*)}$ and $\zeta_m^{(*)}$ are set to be zero, they yield the constraints

$$\lambda_m + \alpha_m = C, \tag{30}$$

$$\eta_m + \beta_m = C,$$
$$\lambda_m^* + \alpha_m^* = C,$$
$$\eta_m^* + \beta_m^* = C,$$

for $\xi_m, \xi_m^* \in [0, rC]$, and the constrains

$$\lambda_m + \alpha_m = \frac{1}{r}\xi_m, \tag{31}$$

$$\eta_m + \beta_m = \frac{1}{r}\zeta_m,$$
$$\lambda_m^* + \alpha_m^* = \frac{1}{r}\zeta_m^*,$$
$$\eta_m^* + \beta_m^* = \frac{1}{r}\zeta_m^*,$$

for $\zeta_m, \zeta_m^* \in [rC, +\infty)$.

Let

$$\mathbf{G}(u, v) = \langle \Phi(z_u), \Phi(z_v) \rangle. \tag{32}$$

Substituting equations (30), (31) and (29) into equation (24) yields the dual optimization problem

$$\max \quad -\frac{1}{2}\Psi^H(\mathbf{G} + r\mathbf{I})\Psi + \mathcal{R}(\Psi^H \mathbf{y}) - (\boldsymbol{\alpha} + \boldsymbol{\alpha}^* + \boldsymbol{\beta} + \boldsymbol{\beta}^*)\mathbf{1}\varepsilon, \tag{33}$$

being constrained to

$$0 \leq \alpha^{(*)}, \beta^{(*)} \leq C, \tag{34}$$

where $\Psi = [\psi_1, \ldots, \psi_n]$, $\mathbf{I}$ and $\mathbf{1}$ are the identity matrix and the all-ones column vector, respectively, $\boldsymbol{\alpha}^{(*)} = [\alpha_1^{(*)}, \ldots, \alpha_n^{(*)}]^T$, $\mathbf{y} = [y_1, \ldots, y_n]^T$.

Note that (33) is a quadratic form and real-valued. It represents a natural extension of the dual functional in SVM real regression for complex-valued problems. Optimizing (33) with respect to $\{\alpha\}$, $\{\alpha^*\}$, $\{\beta\}$, $\{\beta^*\}$, the final solution is expressed as

$$g(z) = \sum_{m=1}^{n} \overline{\psi_m} K(z, z_m) = \sum_{m=1}^{n} \overline{\psi_m} \frac{1}{1 - \overline{z_m} z}. \tag{35}$$

As in the usual SVM framework, by letting $\varepsilon > 0$, we have only a subset of the Lagrange multipliers being nonzero, and thus we obtain the following sparse solution.

$$g(z) = \sum_{m=1}^{n} \overline{\psi_m} \frac{1}{1 - \overline{z_m} z} = \sum_{m \in J} \overline{\psi_m} \frac{1}{1 - \overline{z_m} z}, \tag{36}$$

where $J = \{m : \psi_m \neq 0\}$. Obviously, $|J| \leq n$.

Note that the obtained coefficients $\overline{\psi_m}$ may be complex numbers. In order to obtain an approximation with real coefficients, we use the conjugate poles $\{\overline{z}_m\}$ and coefficients $\{\psi_m\}$ to obtain a function $g^*(z)$, viz,

$$g^*(z) = \sum_{m \in J} \psi_m \frac{1}{1 - z_m z}. \tag{37}$$

For each index m,

$$\psi_m \frac{1}{1 - z_m z} + \overline{\psi_m} \frac{1}{1 - \overline{z_m} z}$$

is a rational function of real coefficients, the sum

$$f_M(z) = \frac{1}{2}(g(z) + g^*(z)) \ (M = 2|J|) \tag{38}$$

is thus a rational function of real coefficients of the degree at most $M = 2|J|$. f_M, in particular, satisfies $\overline{f_M(z)} = f_M(\overline{z})$.

It can be proved that f_M has twice generalization bounds as g, that is

Corollary 3.1 *Let $g \in \mathcal{F}$ be a function satisfying Theorem 2.1. Then there is a constant c such that with probability at least $1 - \delta$ over randomly drawn training sets S of size l and for all γ, $0 < \gamma \leq \theta$, the probability that the function f_M defined by (38) has error larger than θ on a randomly chosen input is bounded by*

$$2\epsilon'(l, \delta, \gamma) = \frac{2c}{l} \left(\frac{4(\|\mathbf{w}\|_2^2 R^2 + (\|\xi\|_1^2 + \|\xi^*\|_1^2) \log(2/\gamma))}{\gamma^2} \log^2 l + 2 \log \frac{2}{\delta} \right), \tag{39}$$

where ξ and ξ^* are defined by g.

Proof. We observe that in the space $X \times \mathbb{C}$,

$$\{(z, y) : |f_M(z) - y| > \theta\} \subseteq \{(z, y) : |g(z) - y| > \theta\} \cup \{(z, y) : |g^*(z) - y| > \theta\}.$$

Denote by $P_{X \times \mathbb{C}}$ the probability on $X \times \mathbb{C}$ and

$$\mathrm{err}_{P_{X \times \mathbb{C}}}(g, \theta) = P_{X \times \mathbb{C}}(\{(z, y) \in X \times \mathbb{C} \ : \ |g(z) - y| > \theta\}).$$

Since

$$|g^*(z) - y| = |\overline{g^*(z) - y}| = |g(\bar{z}) - \bar{y}|,$$

there holds

$$\mathrm{err}_{P_{X\times\mathbb{C}}}(g,\theta) = \mathrm{err}_{P_{X\times\mathbb{C}}}(g^*,\theta).$$

By Theorem 2.1, for some c in the relation (6) there holds, for $|S| = l$,

$$P^l_{X\times\mathbb{C}}(\{S : \mathrm{err}_{P_{X\times\mathbb{C}}}(g,\theta) > \epsilon'_{g,c}(l,\delta,\gamma)\}) < \delta,$$

where the $\epsilon_{g,c}$ is defined as the same as in (8) but here with the dependence of g and c. Since

$$\mathrm{err}_{P_{X\times\mathbb{C}}}(f_M,\theta) \le \mathrm{err}_{P_{X\times\mathbb{C}}}(g,\theta) + \mathrm{err}_{P_{X\times\mathbb{C}}}(g^*,\theta),$$

we have that the set

$$\{(z,y) \in X\times\mathbb{C} \ : \ \mathrm{err}_{P_{X\times\mathbb{C}}}(f_M,\theta) > 2\epsilon'_{g,c}(l,\delta,\gamma)\}$$

is contained in the union of the two sets

$$\begin{aligned}&\{(z,y) \in X\times\mathbb{C} \ : \ \mathrm{err}_{P_{X\times\mathbb{C}}}(g,\theta) > \epsilon'_{g,c}(l,\delta,\gamma)\}\\ &\cup\{(z,y) \in X\times\mathbb{C} \ : \ \mathrm{err}_{P_{X\times\mathbb{C}}}(g^*,\theta) > \epsilon'_{g,c}(l,\delta,\gamma)\}.\end{aligned}$$

Therefore, on randomly chosen training sets S of cardinality l, for a randomly chosen input, we have, in the product space,

$$\begin{aligned}&P^l_{X\times\mathbb{C}}(\{S : \ \mathrm{err}_P(f_M,\theta) > 2\epsilon'_{g,c}(l,\delta,\gamma)\})\\ &\le P^l_{X\times\mathbb{C}}(\{S : \mathrm{err}_{P_{X\times\mathbb{C}}}(g,\theta) > \epsilon'_{g,c}(l,\delta,\gamma)\})\\ &+P^l_{X\times\mathbb{C}}(\{S : \mathrm{err}_{P_{X\times\mathbb{C}}}(g^*,\theta) > \epsilon'_{g,c}(l,\delta,\gamma)\}) < 2\delta.\end{aligned}$$

The desired result then follows. □

Remark 3.5 *Researchers have developed methods to improve efficiency and quality of support vector machine by selecting samples and significant support vectors. For instance, the method given by [6] is to select samples $\{(z_m,y_m)\}_{m=1}^n$ based on the properties of the kernel function. The method proposed by [15] is to select the representative support vectors to obtain a simpler model which avoids the over-fitting problem.*

Remark 3.6 *We have been working on Szegő kernel in the proposed algorithm. In fact, the method is applicable to other complex reproducing kernel. For instance, it is applicable to the Bergman kernel if we want to approximate functions in the Bergman spaces.*

4 EXAMPLE

In this section, we evaluate the performance of the proposed complex SVM algorithm by comparing it with other methods. The results are compared with Core-AFD (Adaptive Fourier Decomposition) method given by [29] in the noise corrupted case and the noiseless case. We adopt the example used in [22, 24], i.e.,

$$f(z) = \frac{z^3(0.0247z + 0.355)}{(1 - 0.9048z)(1 - 0.3679z)}. \tag{40}$$

The function $f(z)$ is assumed to be the true system function which is usually unknown. Given $m = 600$ measurements of (40) in the interval $[0, \pi)$ corresponding to a half circle. We have $N = 2m$ points on the full circle. The frequency responses of FIR model, Laguerre model, Core-AFD method and our method are compared in Figure 1, Figure 2, Figure 3 and Figure 4, respectively. In the noise corrupted case, we assume that the measurements are sampled with added Gaussian noise with SNR= 20. Figure 1 shows the frequency response of 4th order, 7th order and 19th order FIR model, respectively. Even the 19th order FIR model can not give satisfactory result. Figure 2 shows the frequency response of 4th order, 7th order and 10th order Laguerre model with $a = 0.3879, a = 0.9048, a = 0.7165$, respectively. Figure 3 and Figure 4 show the frequency response of 7th Core-AFD and our proposed method in the noise free case and Gaussian noise corrupted case.

In the proposed algorithm, we first get an approximation $\tilde{f}$ of $f(z)$ by using the measurements. The role of $\tilde{f}$ is to assert the data of the training points. Then we choose 600 samples $\{(z_i, y_i)\}$ evenly which satisfy $y_i = \tilde{f}(z_i)$ as training data.

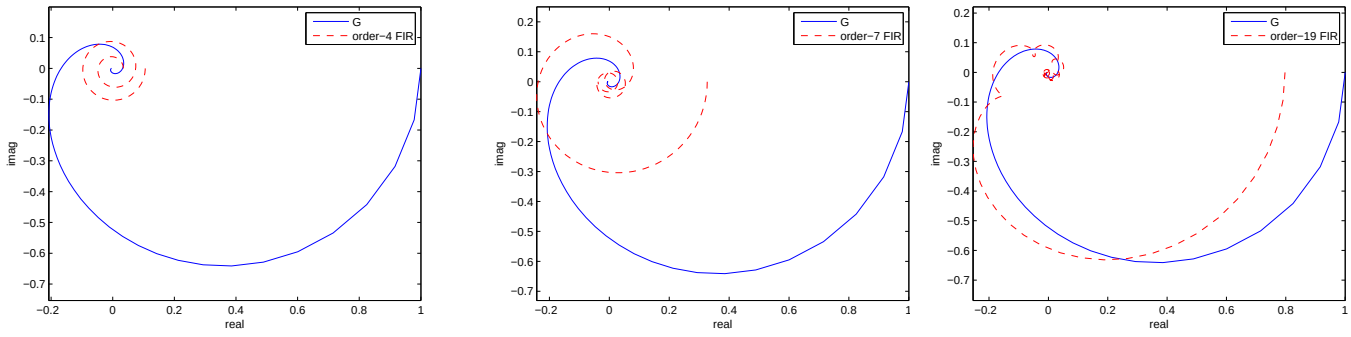

Figure 1: Frequency response of FIR model.

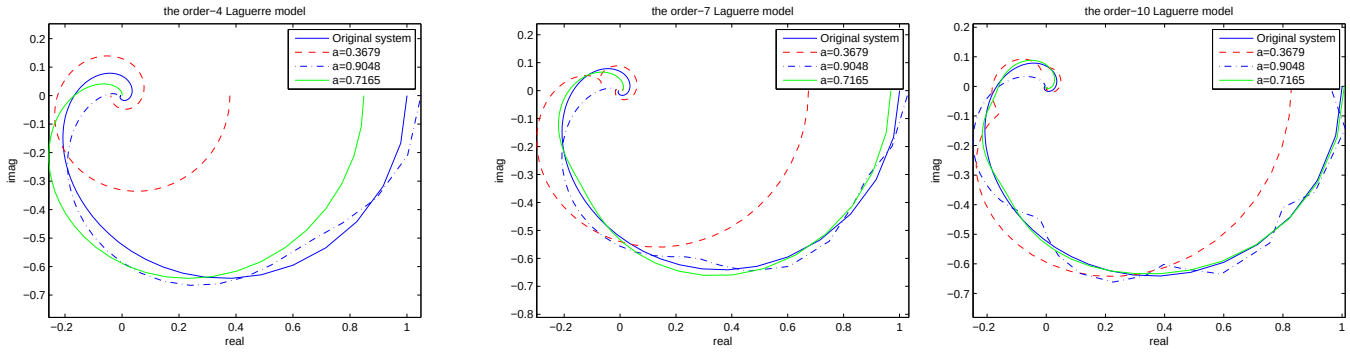

Figure 2: Frequency response of Laguerre model.

In Step 3, the free parameters are explored. According to $C \in (1, 10^2)$, $rC \in (10^{-4}, 0.5)$, $\varepsilon \in (0, 1)$, we set $C = 10, r = 0.1, \varepsilon = 0.002$. After using the proposed algorithm, we use the orthogonal least-squares method proposed by [15] to select the representative support vectors to improve the generalization capability of support vector

	z_1	z_2	z_3	z_4	z_5	z_6	z_7
Without noise	0.9000	0.589	-0.12i	-0.54i	0.98	0.12+0.3i	0.999i
With noise	0.175	0.29	0.4199	0.8999	0.147i	-0.1i	0.12i

Table 1: The selected support vectors in both cases.

machine. We select 7 representative support vectors. Then we obtain rational approximation of $\tilde{f}(z)$. We can see that the proposed method is better than the Core-AFD for this example.

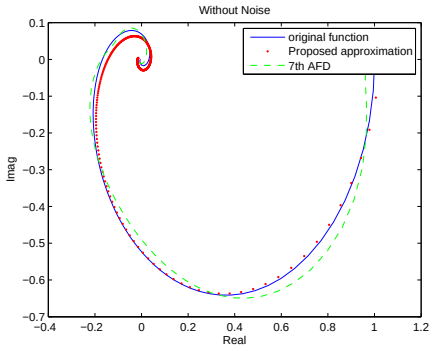

Figure 3: Without noise

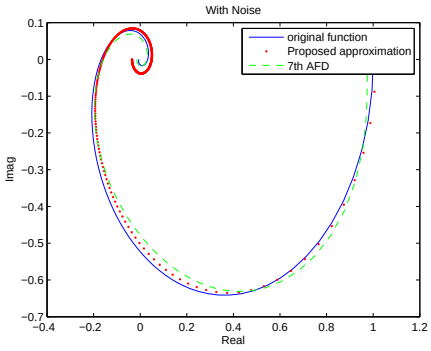

Figure 4: With noise

Remark 4.1 *The reasons of comparison with AFD lay on the facts that they both are of the rational approximation kind; and AFD has been well accepted as an effective method. There are other types of AFD developed by Qian et al., for instance, the Cyclic AFD [38] and Unwending AFD [37]. Further comparisons show that they all have their respective merits. In the noiseless data case, the Cyclic AFD and Unwending AFD have better performance than the proposed algorithm. However, the running time of the proposed algorithm is the shortest. In the noise corrupted data case, the Cyclic AFD does not perform with adequate stability since it is heavily dependent on the initial value. The proposed algorithm has a stable performance on the noise corrupted data. The Unwending AFD is theoretically advanced, but it requires accurate computation of Hilbert transform. This reduces its effectiveness at the present time. The proposed algorithm is better to treat data with noise, which has a generalization bound result available for error estimation and has a shorter running time.*

5 CONCLUSIONS

We discuss generalization bounds for learning complex-valued data which are a theoretical foundation of complex support vector regression machine. A complex support vector machine based on the Szegő kernel is formulated and is subsequently used in the frequency domain identification problem of discrete linear time-invariant system (LTIS). To conclude, compared with existing methods, the newly proposed method has the advantage in terms of stability, fast computation and the generalization bound estimation.

References

[1] H. Akçay and B. Ninness. Orthonormal basis functions for modelling continuous-time systems. *Signal Processing*, 77(3):261 – 274, 1999.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[3] P. Bouboulis and S. Theodoridis. Extension of Wirtinger's Calculus to Reproducing Kernel Hilbert Spaces and the Complex Kernel LMS. *IEEE Transactions on Signal Processing*, 59:964–978, 2011.

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

[5] A. Charrada and A. Samet. Complex support vector machine regression for robust channel estimation in let downlink system. *International Journal of Computer Networks & Communications*, 4, 2012.

[6] Q. Chen and S. Chen. Sample selection algorithm based on kernel function in support vector machine. *Computer Engineering and Design*, 31:2266–2269, 2010.

[7] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[8] D. K. de Vries and P. Van den Hof. Frequency domain identification with generalized orthonormal basis functions. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 2, pages 1240–1245 vol.2, Dec 1995.

[9] H. Drucker, C. J. Burges, L. Kaufman, C. J. C, B. L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines, 1996.

[10] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Willey & Sons, New Yotk, 1973.

[11] M.-G. Garcia, J. Rojo-Alvarez, F. Atienza, and M. Martinez-Ramon. Support vector machines for robust channel estimation in ofdm. *Signal Processing Letters, IEEE*, 13(7):397–400, July 2006.

[12] J. Garnett. *Bounded analytic functions*. Pure and Applied Mathematics. Elsevier Science, 1981.

[13] G. Gu and P. P. Khargonekar. A class of algorithms for identification in h_∞. *Automatica*, 28(2):299 – 312, 1992.

[14] P. Heuberger, P. van den Hof, and B. Wahlberg. *Modelling and Identification with Rational Orthogonal Basis Functions*. Springer, 2005.

[15] W. Lee, C. Yang, and S. J. Lee. Support vector selection for regression machines. *Fifth International Workshop on Computational Intelligence & Applications*, November 2009.

[16] J. Li, K. Miyashita, T. Kato, and S. Miyazaki. Gps time series modeling by autoregressive moving average method: Application to the crustal deformation in central japan. *Earth, Planets and Space*, 52(3):155–162, 2000.

[17] L. Ljung. *System Identification: Theory for the User*. Pearson Education, 1998.

[18] P. M. Mäkilä. Approximation of stable systems by laguerre filters. *Automatica*, 26(2):333 – 345, 1990.

[19] O. L. Mangasarian. *Nonlinear Programming*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1994.

[20] D. Mattera and S. Haykin. Advances in kernel methods. chapter Support Vector Machines for Dynamic Reconstruction of a Chaotic System, pages 211–241. MIT Press, Cambridge, MA, USA, 1999.

[21] G. McCormick. *Nonlinear Programming: Theory, Algorithms and Applications*. A Wiley-Interscience publication. Wiley, 1983.

[22] W. Mi and T. Qian. Frequency-domain identification: An algorithm based on an adaptive rational orthogonal system. *Automatica*, 48(6):1154 – 1162, 2012.

[23] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 511–520, Sep 1997.

[24] B. Ninness and F. Gustafsson. A unifying construction of orthonormal bases for system identification. *Automatic Control, IEEE Transactions on*, 42(4):515–521, Apr 1997.

[25] B. Ninness, H. Hjalmarsson, and F. Gustafsson. The fundamental role of general orthonormal bases in system identification. *Automatic Control, IEEE Transactions on*, 44(7):1384–1406, Jul 1999.

[26] J. Partington. *Interpolation, Identification, and Sampling*. London Mathematical Society monographs. Clarendon Press, 1997.

[27] M. Pontil and A. Verri. Support vector machines for 3-d object recognition. *IEEE Trans. PAMI*, 20:637–646, 1998.

[28] T. QIAN and Y. WANG. Remarks on adaptive fourier decomposition. *International Journal of Wavelets, Multiresolution and Information Processing*, 11(01):1350007, 2013.

[29] T. Qian and Y.-B. Wang. Adaptive fourier series-a variation of greedy algorithm. *Advances in Computational Mathematics*, 34(3):279–293, 2011.

[30] J. Rojo-álvarez, G. Camps-Valls, M. Martinez-Ramón, E. Soria-Olivas, A. Návia-Vazquez, and A. Figueiras-Vidal. Support vector machines framework for linear signal processing. *Signal Processing*, 85(12):2316 – 2326, 2005.

[31] B. Scholkopf, S. Mika, C. Burges, P. Knirsch, K. Muller, G. Ratsch, and A. Smola. Input space versus feature space in kernel-based methods. *Neural Networks, IEEE Transactions on*, 10(5):1000–1017, Sep 1999.

[32] B. Scholkopf and K. Sung. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Processing*, 45:2758C2765, 1997.

[33] J. Shawe-taylor and N. Cristianini. On the generalisation of soft margin algorithms. *IEEE Transactions on Information Theory*, 48:2721–2735, 2002.

[34] A. Smola and B. Schlköpf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004.

[35] M. Smola, A. J. Smola, G. Ratsch, B. Scholkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines, 1997.

[36] O. Szász. On closed sets of rational functions. *Annali di Matematica Pura ed Applicata*, 34(1):195–218, 1953.

[37] T.Qian. Intrinsic mono-component decomposition of functions: An advance of fourier theory. *Mathematical Methods in the Applied Sciences*, 33:880C891, 2010.

[38] T.Qian. Cyclic afd algorithm for best rational approximation. *Mathematical Methods in the Applied Sciences*, 37:846C859, 2014.

[39] V. N. Vapnik. The nature of statistical learning theory, 1995.

[40] A. C. V.Vapnik. Necessary and sufficient conditions for consistency in the empirical risk minimization method. *pattern Recognition and Image Analysis*, 3:283–305, 1991.

[41] B. Wahlberg. Identification of resonant systems using kautz filters. In *Decision and Control, 1991., Proceedings of the 30th IEEE Conference on*, pages 2005–2010 vol.2, Dec 1991.

[42] B. Wahlberg. System identification using laguerre models. *Automatic Control, IEEE Transactions on*, 36(5):551–562, May 1991.

[43] B. Wahlberg and L. Lennart. Design variables for bias distribution in transfer function estimation. *IEEE Transactions on Automatic Control*, 31:133–144, 1986.