Report: Complex Support Vector Detector for Large MIMO System

Tianpei Chen

Department of Electrical and Computer Engineering

McGill University

October 6, 2015

I. Introduction

One of the biggest challenges the researchers and industry practitioners are facing in wireless communication area is how to bridge the sharp gap between increasing demand of high speed communication of rich multimedia information with high level Quality of Service (QoS) requirements and the limited radio frequency spectrum over a complex space-time varying environment. The most promising technology for solving this problem, Multiple Input Multiple Output (MIMO) technology has been of immense research interest over the last several tens of years is incorporated into the emerging wireless broadband standard like 802.11ac [1] long-term evolution (LTE) [2]. The core idea of MIMO system is to use multiple antennas at both transmitting and receiving end, so that multiplexing gain (multiple parallel spatial data pipelines that can improve bandwidth efficiency) and diversity gain (better reliability of communication link) is obtained by exploiting the spatial domain. Large MIMO (also called Massive MIMO) is an upgraded version of conventional MIMO technology employing hundreds of low power low price antennas at base station (BS), that serves several tens of terminals simultaneously. This technology can achieve full potential of conventional MIMO system while providing additional power efficiency as well as system robustness both to unintended man-made interference and intentional jamming. [3] [4].

The price paid for large MIMO system is the increased complexities for signal processing at both transmitting and receiving end. The uplink Detector is one of the key components in large MIMO systems. With orders magnitude more antennas at the BS, benefits and challenges coexist in designing of detection algorithms for the uplink communication of large MIMO systems. On one hand, a large number of receive antennas provide potential of large diversity gains, on the other hand, complexity of the algorithm becomes crucial to make the system practical.

Vertical Bell Laboratories Layered Space-Time (V-BLAST) architecture for MIMO system can achieve high spectrum efficiency by spatial multiplexing (SM), that is, each transmit antenna transmits independent symbol streams. However the optimal maximum likelihood detector (MLD) for V-BLAST systems that perform exhaustive search has a complexity that increases exponentially with number of transmitted antennas, which is prohibitive for practical applications.

As alternatives to MLD, linear detectors (LD) such as zero-forcing (ZF) and minimum mean square error (MMSE) with optimized ordering sequential interference cancellation (ZF-OSIC, MMSE-OSIC) are exploited in V-BLAST architecture [5] [6] [7], however the performance of ZF-OSIC and MMMSE-OSIC are inferior comparing to MLD.

Sphere Decoder (SD) [8] is the most prominent algorithm that utilizes the lattice structure of

MIMO systems, which can achieve optimal performance with relatively much lower complexity comparing to MLD. However, SD has two major shortages that make it problematic to be integrated into a practical systems. The first shortage is SD has various complexities under different signal to noise ratios (SNR), while a constant processing data rate is required for hardware. The second shortage is SD's complexity still has a lower bound for complexity that increases exponentially with the number of transmit antennas and the order of modulation scheme [9]. The fixed complexity sphere decoder (FCSD) [10] makes it possible to achieve near optimal performance with a fixed complexity under different value of SNR. The FCSD inherits the principle of list based searching algorithms, which first generate a list of candidate symbol vectors and then the best candidate is chosen as the solution. The other sub optimal detectors belong to this class include Generalized Parallel Interference Cancellation (GPIC) [11] and Selection based MMSE-OSIC(sel-MMSE-OSIC) [12]. However, all these list based searching algorithms have the same shortage - their complexities increase exponentially with the number of transmit antennas and the order of modulation scheme [12]. Therefore, such algorithms are prohibitive when it comes to a large number of antennas or a high order modulation scheme, for example in IEEE 802.11ac standard [1], the modulation scheme is 256QAM.

Besides the above detection algorithms designed for conventional MIMO systems, in the last several years, a set of detection algorithms have been proposed for large MIMO systems with complexities that are comparable with MMSE detector and near-optimal performance. such algorithms include likelihood ascend searching (LAS) algorithms [13] [14], Tabu search based algorithms which have superior performance compared to LAS detectors because local minima can be avoided (e.g. Layered Tabu search (LTS) [15], Random Restart Reactive Tabu search (R3TS) [16]), Message passing technique based algorithms (e.g. Belief propagation (BP) detectors based on graphic model and Gaussian Approximation (GA) [17] [18] [19] [20]), Probabilistic Data Association based algorithms [21], Monte Carlo sampling based algorithms (e.g. Markov Chain Monte Carlo (MCMC) algorithm [22]) and Lattice Reduction (LR) aided algorithms [23].

Firmly grounded in framework of statistical learning theory, the Support Vector Machine (SVM) technique has become a powerful tool to solve real world supervised learning problems such as classification, regression and prediction. the SVM method is a nonlinear generalization of Generalized Portrait algorithm developed by Vapnik in 1960s [24] [25], which can provide good generalization performance [26].

Interest in SVM boosted since 1990s, promoted by the works of Vapnik and co-workers at AT& T Bell laboratory [27] [28] [29] [30] [31] [32]. Moreover, the kernel based methods [26] solve nonlinear learning tasks by mapping input data sets into high dimensional feature spaces, and replacing inner products of feature mappings by computational inexpensive kernel functions discarding the actual structure of the feature space. This rationale is supported mathematically by the notion of Reproducing Kernel Hilbert Space (RKHS). Based on the same regularized risk function principle, ϵ-Support Vector Regression (ϵ-SVR) was developed [29] [33].

Similar to SVM, the ϵ-SVR solves an original optimization problem by transforming it into a Lagrange dual optimization problem, which can be solved by Quadratic Programming (QP). Sequential Minimal Optimization (SMO) algorithm was proposed as a fast algorithm to solve this QP problem by decomposing the it into sub QP problems and solving them analytically [34]. Therefore, the computational intensive numerical method can be avoided. A more general method is decomposition solver, which refers to a set of algorithms that separate the optimization variables (Lagrange multipliers) into two sets W and N, W is the work set and N contains the remaining optimization variables. In each iteration, only the optimization variables in the work set is optimized while keeping other variables fixed. The SMO algorithm is an extreme case of decomposition solver. An important issue of decomposition solver is the choice of the work set. One strategy is to choose Karush-Kuhn-Tucker (KKT) condition violators, ensuring final converge [35]. Because of the linear constraint inducted by offset, the SMO algorithm restricts the size of the work set to 2. In [36], a method to train SVM without offset was proposed, with the comparable performance to the SVM with offset. A set of sequential single variable work set selection strategies, which require $O(n)$ searching time are proposed. The optimal double variable work set selection strategy, which performs exhaustively searching, however, requires $O(n^2)$ searching time. The authors demonstrate that with the combination of two proposed single variable work set selection strategies, convergence can be achieved by a iteration time that is as few as optimal double variable work set selection strategy.

The mathematical foundation of kernel based methods is RKHS which is defined in complex domain, however most of the practitioners are dealing with real data sets. In communication and signal processing area, the channel gains, signals, waveforms etc. are all represented in complex form. Recently, a pure complex SVR & SVM based on complex kernel was proposed in [37], which can deal with the complex data set purely in complex domain. The results in [37] demonstrate better performances as well as reduced complexity comparing to simply split learning task into two real case by real kernel. Based on this work, we derive of a complexity-performance controllable detector for large MIMO systems based on a dual channel complex

SVR (CSVR). The detector can work in two parallel real SVR channels which can be solved independently. Moreover, only the real part of kernel matrix is needed in both channels. This means a large amount of computation can be reduced. Based on the discrete time MIMO channel model, in our regression model, this CSVR-detector is constructed without offset, Therefore, for each real SVR without offset, in principle, only one variable is needed to be updated in each iteration, In our scheme, a sequential single variable selection strategy is proposed. By this strategy, two variables can be updated at each iteration, with much smaller searching time.

II. Brief Introduction to ϵ-Support Vector Regression

A. Regression Model

Suppose we are given training data set $((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_L, y_L))$, L denotes the number of training samples, $\mathbf{x}_i \in \mathbb{R}^V$ denotes input data vector, V is the number of features in $\mathbf{x}_i$. y_i denotes output. Let $\mathbf{w}$ denotes regression coefficient vector, $\Phi(\mathbf{x}_i)$ denotes the mapping of $\mathbf{x}_i$ to higher dimensional feature space, $\mathbf{w}, \Phi(\mathbf{x}_i) \in \mathbb{R}^\Lambda$, $\Lambda \in \mathbb{R}$ denotes the dimension of mapped feature space (For linear model, $\mathbf{x}_i = \Phi(\mathbf{x}_i)$). The regression model (either linear or non-linear) is given by

$$y_i = \mathbf{w}^T \Phi(\mathbf{x}_i) + b \quad i = 1 \cdots L \tag{1}$$

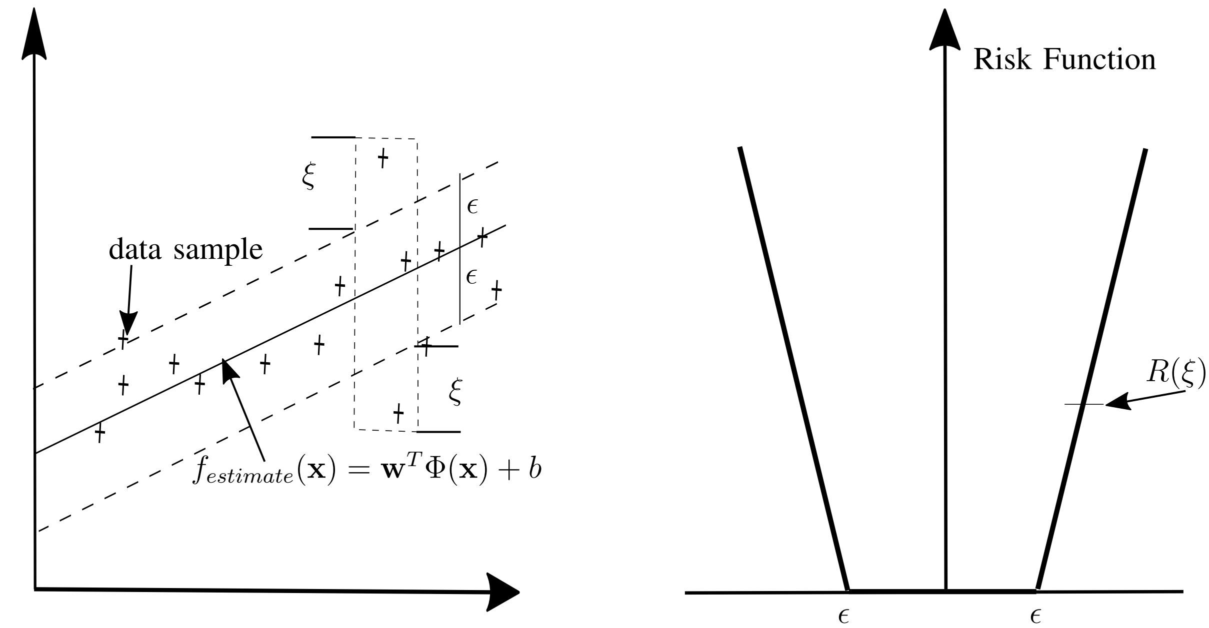

Fig. 1. ϵ-Support Vector Regression and Risk Functional

We present the primal optimization problem directly

$$\min_{\mathbf{w}} f(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{L}(R(\xi_i) + R(\hat{\xi}_i))$$

$$s.t. \begin{cases} y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b \leq \epsilon + \xi_i, i = 1, 2\cdots, L \\ \mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i \leq \epsilon + \hat{\xi}_i, i = 1, 2\cdots, L \\ \epsilon_i, \xi_i, \hat{\xi}_i \geq 0, i = 1, 2\cdots, L \end{cases} \tag{2}$$

In (2), $\frac{1}{2}||\mathbf{w}||^2$ is the regularization term in order to ensure the flatness of regression model, ϵ

denotes the precision, As shown in Fig 1, the area between two dash line is called ϵ tube. Only those data samples located outside ϵ tube contribute to cost. Furthermore ξ_i and $\hat{\xi}_i$ denote slack variables that cope with noise of input data samples, $R(u)$ denotes risk function, the simplest risk function is $R(u) = u$, The type of cost function is determined by the statistical distribution of noise [33]. For example if the noise is Gaussian noise, then the optimal cost function is $R(u) = \frac{1}{2}u^2$. The term $C\sum_{i=1}^{L}(R(\xi_i) + R(\hat{\xi}_i))$ denotes the penalty of noise, $C \in \mathbb{R}$ and $C \geq 0$ that controls the trade off between regularization term and cost function term.

B. Cost Function

The optimal cost function in (2) can be derived based on maximum likelihood (ML) principle. Assume the data samples $\mathbf{x}_i$ in data set are iid, Let $f_{true}(\mathbf{x}_i), i = 1, 2, \cdots L$ denotes true regression function. the underlying assumption is $y_i = f_{true}(\mathbf{x}_i) + \xi_i, i = 1, 2, \cdots, L$, ξ_i denotes additive noise of the ith data sample, with probability density function (pdf) $Pr(\cdot)$. Let $P(\cdot)$ denotes the pdf of y_i. Based on ML principle we want to

$$\max_{f} \quad \prod_{i=1}^{L} P(y_i|f(\mathbf{x}_i)) = \prod_{i=1}^{L} P(f(\mathbf{x}_i) + \xi_i|f(\mathbf{x}_i)) = \prod_{i=1}^{L} Pr(\xi_i) = \prod_{i=1}^{L} Pr(y_i - f(\mathbf{x}_i)) \tag{3}$$

Take the logarithm of $\prod_{i=1}^{L} Pr(y_i - f(\mathbf{x}_i))$, we have

$$\sum_{i=1}^{L} log(Pr(y_i - f(\mathbf{x}_i))), \tag{4}$$

maximizing (4) is equivalent to minimizing $-\sum_{i=1}^{L} log(Pr(y_i - f(\mathbf{x}_i)))$ Let $c(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$ denotes the ith cost function, which is defined as

$$c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) = -log(P(y_i - f(\mathbf{x}_i))). \tag{5}$$

Thus (3) can be rewritten as

$$\min_{f} \quad \sum_{i=1}^{L} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)), \tag{6}$$

In ϵ-SVR, in order to provide more flexibility to precision control, Vapnik's ϵ-insensitive function, as shown in (7), is applied to (5).

$$|u|_\epsilon = \begin{cases} 0 & if \quad |u| < \epsilon \\ |u| - \epsilon & otherwise \end{cases} \tag{7}$$

Thus the final form of cost function in ϵ-SVR can be written as

$$\tilde{c}(y_i, \mathbf{x}_i, f(\mathbf{x}_i)) = -log(Pr(|y_i - f(\mathbf{x}_i)|_\epsilon)), \tag{8}$$

Consider the cost function term in (2),

$$\sum_{i=1}^{L} R(\xi_i) + R(\hat{\xi}_i) = \sum_{i=1}^{L} \tilde{c}(y_i, \mathbf{x}_i, f(\mathbf{x}_i)) = \sum_{i=1}^{L} -log(Pr(|y_i - f(\mathbf{x}_i)|_\epsilon)) \tag{9}$$

the cost function term is determined according to noise distribution.

C. Lagrange Duality

According to Lagrange Theorem, the constraint optimization problem (2) can be transformed to Lagrangian dual form by combining the original optimization function with inequality constraints, the combination coefficient is called Lagrange multiplier. The Lagrange function is given by.

$$
\begin{aligned}
&L(\mathbf{w}, b, \xi_i, \hat{\xi}_i, \alpha_i, \hat{\alpha}_i, \eta_i, \hat{\eta}_i) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{L}(R(\xi_i) + R(\hat{\xi}_i)) - \sum_{i=1}^{L}(\eta_i\xi_i + \hat{\eta}_i\hat{\xi}_i) \\
&+\sum_{i=1}^{L}\alpha_i(y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \sum_{i=1}^{L}\hat{\alpha}_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i) \\
&s.t. \begin{cases} \eta_i, \hat{\eta}_i, \alpha_i, \hat{\alpha}_i \geq 0, i = 1, 2, \cdots L \\ \xi_i, \hat{\xi}_i \geq 0, i = 1, 2, \cdots L \end{cases}
\end{aligned} \tag{10}
$$

where η_i, $\hat{\eta}_i$, α_i, $\hat{\alpha}_i$ are Lagrange multipliers.

The sufficient and necessary conditions such that a solution $\mathbf{w}$ of the constrained optimization problem in (2) satisfies, are called Karush-Kuhn-Tucker (KKT) conditions. Here we briefly introduce how the dual objective problem is derived from KKT conditions.

Assume a constraint optimization problem is given by

$$
\begin{aligned}
&\min_{\mathbf{w}} \quad f(\mathbf{w}) \\
&s.t. \quad c_i(\mathbf{w}) \leq 0, i = 1, 2, \ldots L,
\end{aligned} \tag{11}
$$

its Lagrange function is given by

$$
L(\mathbf{w}, a_i) = f(\mathbf{w}) + \sum_{i=1}^{L} a_i c_i(\mathbf{w}), \tag{12}
$$

where a_i denote Lagrange multipliers. Based on Theorem 6.21 in [26], For a variable set $[\bar{\mathbf{w}}, \bar{a}_i]$, $\bar{\mathbf{w}}$ is the solution to (11) only when the following inequalities are satisfied

$$
L(\mathbf{x}, \bar{a}_i) \geq L(\bar{\mathbf{x}}, \bar{a}_i) \geq L(\bar{\mathbf{w}}, a_i) \tag{13}
$$

This inequalities yield KKT conditions (see Theorem 6.26 [26]), which are

$$\partial_{\mathbf{w}} L(\bar{\mathbf{w}}, a_i) = \partial_{\mathbf{w}} f(\bar{\mathbf{w}}) + \sum_{i=1}^{L} a_i \partial_{\mathbf{w}} c_i(\bar{\mathbf{w}}) = 0, \tag{14}$$

$$\partial_{a_i} L(\bar{\mathbf{w}}, \bar{a}_i) = c_i(\bar{\mathbf{w}}) \leq 0, i = 1, 2, \ldots L \tag{15}$$

$$\bar{a}_i c_i(\bar{\mathbf{w}}) = 0, i = 1, 2, \ldots L \tag{16}$$

In order to satisfy the first inequality in (13), (14) has to hold, applying (14) to (10), which are

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i) \Phi(\mathbf{x}_i) = 0 \tag{17}$$

$$\frac{\partial L}{\partial \xi_i} = C_i R^{'}(\xi_i) - \eta_i - \alpha_i = 0, i = 1, 2, \cdots L \tag{18}$$

$$\frac{\partial L}{\partial \hat{\xi}_i} = C_i R^{'}(\hat{\xi}_i) - \hat{\eta}_i - \hat{\alpha}_i = 0, i = 1, 2, \cdots L \tag{19}$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i) = 0 \tag{20}$$

$$\tag{21}$$

Then by substituting (17)-(20) to (10), (10) can be rewritten as :

$$\theta(\alpha_i, \hat{\alpha}_i) = \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} (\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j) \Phi^T(\mathbf{x}_j) \Phi(\mathbf{x}_i) + C \sum_{i=1}^{L} [R(\xi_i) + R(\hat{\xi}_i)] - \sum_{i=1}^{L} [(CR^{'}(\xi_i) - \alpha_i)\xi_i$$
$$+ (CR^{'}(\hat{\xi}_i) - \hat{\alpha}_i)\hat{\xi}_i] + \sum_{i=1}^{L} \alpha_i (y_i - \sum_{j=1}^{L} (\alpha_j - \hat{\alpha}_j) \Phi^T(\mathbf{x}_j) \Phi(\mathbf{x}_i) - b - \epsilon - \xi_i) +$$
$$\sum_{i=1}^{L} \hat{\alpha}_i (\sum_{j=1}^{L} (\alpha_j - \hat{\alpha}_j) \Phi^T(\mathbf{x}_j) \Phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i) \tag{22}$$

notice in (20), $\sum_{i=1}^{L} (\alpha_i - \hat{\alpha}_i) = 0$, define $\tilde{R}(u) = R(u) - uR^{'}(u)$, (22) can be further simplified to

$$\theta(\alpha_i,\hat{\alpha}_i) = -\frac{1}{2}\sum_{i=1}^{L}\sum_{i=1}^{L}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i) + C\sum_{i=1}^{L}[(\tilde{R}(\xi_i)+\tilde{R}(\hat{\xi}_i)]$$

$$+\sum_{i=1}^{L}[(\alpha_i-\hat{\alpha}_i)y_i-(\alpha_i+\hat{\alpha}_i)\epsilon]$$

$$s.t.\begin{cases}\sum_{i=1}^{l}(\alpha_i-\hat{\alpha}_i)=0\\ 0<\alpha<C\tilde{R}'(\alpha)\\ 0<\hat{\alpha}<C\tilde{R}'(\hat{\alpha})\end{cases} \tag{23}$$

In order to satisfy the second inequality in (13), (15) and (16) have to hold. Satisfying the condition (15) is equivalent to find the maximum of $L(\bar{\mathbf{w}},a_i)$, notice $\theta(\alpha_i,\hat{\alpha}_i)$ is equivalent to $L(\bar{\mathbf{w}},a_i)$ in (13), thus yield the dual optimization problem which is given by

$$\max_{\alpha_i,\hat{\alpha}_i}\quad \theta(\alpha_i,\hat{\alpha}_i) = -\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\Phi^T(\mathbf{x}_j)\Phi(\mathbf{x}_i) + C\sum_{i=1}^{L}(\tilde{R}(\xi_i)+\tilde{R}(\hat{\xi}_i))$$

$$+\sum_{i=i}^{L}[(\alpha_i-\hat{\alpha}_i)y_i-(\alpha_i+\hat{\alpha}_i)\epsilon] \tag{24}$$

Define $\mathbf{a}=[\alpha_1,\alpha_2,\ldots,\alpha_L]^T$, $\hat{\mathbf{a}}=[\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_L]^T$, $\mathbf{y}=[y_1,y_2,\ldots y_L]^T$, $\mathbf{e}=[1,1,\ldots,1]^T\in\mathbb{R}^L$, $\mathbf{e}_i$ denotes the vector that only ith component is 1 while the rest are all 0, $\mathbf{R}_\xi=[\tilde{R}(\xi_1),\tilde{R}(\xi_2),\ldots,\tilde{R}(\xi_L)]^T$, $\mathbf{R}_{\hat{\xi}}=[\tilde{R}(\hat{\xi}_1),\tilde{R}(\hat{\xi}_2),\ldots,\tilde{R}(\hat{\xi}_L)]^T$, $[\mathbf{K}]_{ij}=\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i)$ denotes a component of data kernel matrix at ith row and jth column. An alternative vector form of (24) can be written as

$$\max_{\mathbf{a},\hat{\mathbf{a}}}\theta(\mathbf{a},\hat{\mathbf{a}}) = -\frac{1}{2}(\mathbf{a}-\hat{\mathbf{a}})^T\mathbf{K}(\mathbf{a}-\hat{\mathbf{a}})+(\mathbf{y}-\epsilon\mathbf{e})^T\mathbf{a}+(-\mathbf{y}-\epsilon\mathbf{e})^T\hat{\mathbf{a}}+C\mathbf{e}^T(\mathbf{R}_\xi+\mathbf{R}_{\hat{\xi}}), \tag{25}$$

We define the following $2L$ vectors $\mathbf{a}^{(*)}=\left[\begin{smallmatrix}\mathbf{a}\\ \hat{\mathbf{a}}\end{smallmatrix}\right]$, $\mathbf{v}\in\mathbb{R}^{2L}$,

$$[\mathbf{v}]_i=\begin{cases}1 & i=1,\dots,l\\ -1 & i=l+1,\dots,2l\end{cases} \tag{26}$$

(25) can also be reformulate as

$$\max_{\mathbf{a}^*}\quad \theta(\mathbf{a}^*)=-\frac{1}{2}(\mathbf{a}^*)^T\left[\begin{smallmatrix}\mathbf{K} & -\mathbf{K}\\ -\mathbf{K} & \mathbf{K}\end{smallmatrix}\right]\mathbf{a}^{(*)}+[(\mathbf{y}-\epsilon)^T,(-\mathbf{y}-\epsilon)^T]\mathbf{a}^{(*)}+C\mathbf{e}^T(\mathbf{R}_\xi+\mathbf{R}_{\hat{\xi}}), \tag{27}$$

Condition in (16) is called KKT complementary condition, the value of $\sum_{i=1}^{L}\bar{a}_i c_i(\bar{\mathbf{w}})$ can monitor how close the solution is to the global optimum. Thus it can be used as a stopping criterion. In the constraint optimization problem of (10), the KKT complementary conditions are given by

$$\begin{cases}\alpha_i(y_i-\mathbf{w}^T\Phi(\mathbf{x}_i)-b-\epsilon-\xi_i)=0, i=1,2\cdots L\\ \hat{\alpha}_i(\mathbf{w}^T\Phi(\mathbf{x}_i)+b-y_i-\epsilon-\hat{\xi}_i)=0, i=1,2\cdots L\end{cases} \tag{28}$$

III. System Model

Consider a uncoded complex large MIMO uplink spatial multiplexing (SM) system with N_t users, where each is equipped with transmit antenna. The number of receive antennas at the Base Station (BS) is N_r, $N_r\geq N_t$. Typically large MIMO systems have hundreds of antennas at the BS, as shown in Fig 2.

Bit sequences, which are modulated to complex symbols, are transmitted by the users over a flat fading channel. The discrete time model of the system is given by:

$$\mathbf{y}=\mathbf{H}\mathbf{s}+\mathbf{n}, \tag{29}$$

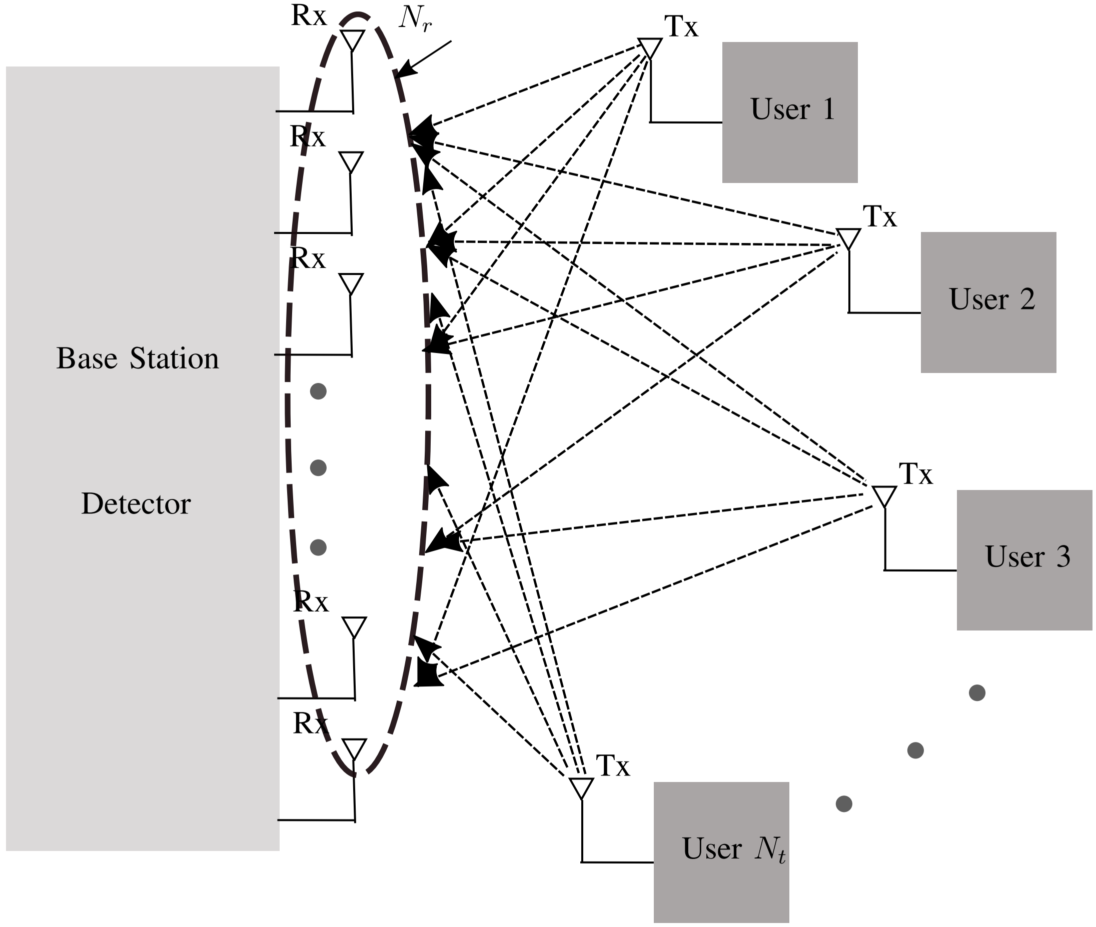

Fig. 2. Large MIMO uplink system

where $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$ is the received symbol vector, $\mathbf{s} \in \mathbb{C}^{N_t}$ is the transmitted symbol vector, with components that are mutually independent and taken from a finite signal constellation alphabet $\mathbb{O}$ (e.g. BPSK, 4-QAM, 16-QAM, 64-QAM), $|\mathbb{O}| = M$. The transmitted symbol vectors $\mathbf{s} \in \mathbb{O}^{N_t}$, satisfy $\mathbb{E}[\mathbf{s}\mathbf{s}^H] = \mathbf{I}_{N_t} E_s$, where E_s denotes the symbol average energy, $\mathbb{E}[\cdot]$ denotes the expectation operation, $\mathbf{I}_{N_t}$ denotes identity matrix of size $N_r \times N_t$. Furthermore $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ denotes the Rayleigh fading channel propagation matrix, each component is independent identically distributed (i.i.d) circularly symmetric complex Gaussian random variable with zero mean and unit variance. Finally, $\mathbf{n} \in \mathbb{C}^{N_r}$ is the additive white Gaussian noise (AWGN) vector with zero mean components and $\mathbb{E}[\mathbf{n}\mathbf{n}^H] = \mathbf{I}_{N_r} N_0$, where N_0 denotes the noise power spectrum density, and hence $\frac{E_s}{N_0}$ is the signal to noise ratio (SNR).

The task of a MIMO detector is to estimate the transmit symbol vector $\mathbf{s}$, based on the knowledge of receive symbol vector $\mathbf{y}$ and channel matrix $\mathbf{H}$.

IV. Dual Channel Complex Support Vector detection for Large MIMO system

Based on discrete time model of large MIMO uplink system in (34), in our regression model, the training data sample at detector is $(\mathbf{h}_1, y_1)(\mathbf{h}_2, y_2), \dots, (\mathbf{h}_{N_r}, y_{N_r})$, where $\mathbf{h}_i$ denotes ith row of channel propagation matrix $\mathbf{H}$, this yields a regression task without offset b:

$$y_i = f_{true}(\mathbf{h}_i) + n, \tag{30}$$

$$f_{true}(\mathbf{h}_i) = \mathbf{h}_i \mathbf{s}, \tag{31}$$

$$\tag{32}$$

where $f_{true}()$ denotes the underlying true function, n denotes additive noise. In this regression problem, receive symbol y is the output data, $\mathbf{h}$ is input data sample, transmitted symbol vector s is regression coefficients. Because the large MIMO system we consider here is complex, we employ complex support vector regression (CSVR) without offset term b. As shown in section II, in order to derive Lagrange duality optimization formula, partial derivatives of objective function with respect to $\mathbf{w}$ and ξ are needed to be calculated, in CSVR, that means take partial derivatives to risk functions which are defined in complex domain. Recently mathematical results of Wirtinger's calculus in Reproducing Kernel Hilbert Space (RKHS) is employed to solve this problem [39]. First we generalize our regression model by complex RKHS, Let $<,>_H$ denotes inner product operation in real RKHS. $<,>_{\mathbb{H}}$ denotes inner products operation in complex RKHS. Assume $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, j, $k \in \mathbb{C}$, complex Hilbert space has the following properties

Property 1. $<\mathbf{x},\mathbf{y}>_{\mathbb{H}}=<\overline{\mathbf{y},\mathbf{x}}>_{\mathbb{H}}$

Property 2. $<j\mathbf{x}+k\mathbf{y},\mathbf{z}>_{\mathbb{H}}=j<\mathbf{x},\mathbf{z}>_{\mathbb{H}}+k<\mathbf{y},\mathbf{z}>_{\mathbb{H}}$

Property 3. $<\mathbf{z},j\mathbf{x}+k\mathbf{y}>=\bar{j}<\mathbf{z},\mathbf{x}>_{\mathbb{H}}+\bar{k}<\mathbf{z},\mathbf{y}>_{\mathbb{H}}$

Lemma 1. $\mathbf{h}_i\mathbf{s}\in<\mathbf{h}_i,\mathbf{s}^*>_{\mathbb{H}}$

Proof. Assume $\mathbf{a}$, $\mathbf{b}\in\mathbb{R}^v$, it can be easily proved

$$\mathbf{a}^T\mathbf{b}\in<\mathbf{a},\mathbf{b}>_H, \tag{33}$$

From Property 1 and Property 3, it is obvious

$$<\mathbf{g},\mathbf{h}>_{\mathbb{H}}=<\mathbf{g}^r,\mathbf{h}^r>_H+<\mathbf{g}^i,\mathbf{h}^i>_H+i(<\mathbf{g}^i,\mathbf{h}^r>_H-<\mathbf{g}^r,\mathbf{h}^i>_H) \tag{34}$$

where $\mathbf{g}$,$\mathbf{h}\in\mathbb{C}^v$, and $\mathbf{g}=\mathbf{g}^r+i\mathbf{g}^i$, $\mathbf{h}=\mathbf{h}^r+i\mathbf{h}^i$. Therefore,

$$\begin{aligned}<\mathbf{h},\mathbf{s}^*>_{\mathbb{H}} &= <\mathbf{h}^r,(\mathbf{s}^*)^r>_H+<\mathbf{h}^i,(\mathbf{s}^*)^i>_H+i(<\mathbf{h}^i,(\mathbf{s}^*)^r>_H-<\mathbf{h}^r,(\mathbf{s}^*)^i>_H)\\ &= <\mathbf{h}^r,\mathbf{s}^r>_H-<\mathbf{h}^i,\mathbf{s}^i>_H+i(<\mathbf{h}^i,\mathbf{s}^r>_H+<\mathbf{h}^r,\mathbf{s}^i>_H),\end{aligned} \tag{35}$$

$$\mathbf{h}\mathbf{s}=\mathbf{h}^r\mathbf{s}^r-\mathbf{h}^i\mathbf{s}^i+i(\mathbf{h}^i\mathbf{s}^r+\mathbf{h}^r\mathbf{s}^i), \tag{36}$$

Because of (38), (40) and (41), $\mathbf{h}_i\mathbf{s} \in < \mathbf{h}_i, \mathbf{s}^* >_{\mathbb{H}}$. □

represent $\mathbf{s}^*$ by $\mathbf{w}$, The general regularized risk function of large MIMO detection in complex RKHS can be formulated:

$$minimize \quad \frac{1}{2}||w||^2_{\mathbb{H}} + C\sum_{k=1}^{N_r}[(R(\xi_k^r) + R(\hat{\xi}_k^r) + R(\xi_k^i) + R(\hat{\xi}_k^i))]$$

$$s.t. \begin{cases} Re(y_k - < \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) \leq \epsilon + \xi_k^r \\ Re(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}} - y_k) \leq \epsilon + \hat{\xi}_k^r \\ Im(y_k - < \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) \leq \epsilon + \xi_k^i \\ Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}} - y_k) \leq \epsilon + \hat{\xi}_k^i \\ \xi^r, \hat{\xi}^r, \xi^i, \hat{\xi}^i \geq 0 \end{cases} \tag{37}$$

where $Re()$ and $Im()$ denote real part and imaginary part of a complex variable, restrictions are set to real and imaginary part of regression function separately. Let $\mathbf{K} = \mathbf{H}\mathbf{H}^H$ denotes the kernel function, $\mathbf{K} = \mathbf{K}^r + i\mathbf{K}^i$, $\mathbf{K}^r$ and $\mathbf{K}^i$ denote matrix of corresponding real part and imaginary part. Similar to the Lagrange duality rational in section II-C, Lagrange function is formulated for (42)

$$
\begin{aligned}
\theta = {} & \frac{1}{2}||w||_{\mathbb{H}}^2 + C\sum_{k=1}^{N_r}[(R(\xi_k^r) + R(\hat{\xi}_k^r) + R(\xi_k^i) + R(\hat{\xi}_k^i))] - \sum_{k=1}^{N_r}(\eta_i\xi_k^r + \hat{\eta}_k\hat{\xi}_k^r + \tau_k\xi_k^i + \hat{\tau}_k\hat{\xi}_k^i) \\
& - \sum_{k=1}^{N_r}\alpha_k(\epsilon + \xi_k^r - Re(y_k) + Re(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}})) - \sum_{k=1}^{N_r}\hat{\alpha}_k(\epsilon + \hat{\xi}_k^r + Re(y_k) - Re(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}})) \\
& - \sum_{k=1}^{N_r}\beta_k(\epsilon + \xi_k^i - Im(y_k) + Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}})) - \sum_{k=1}^{N_r}\hat{\beta}_k(\epsilon + \hat{\xi}_k^i + Im(y_k) - Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}})) \\
& s.t. \begin{cases} \eta, \hat{\eta}, \tau, \hat{\tau}\alpha, \hat{\alpha}, \beta, \hat{\beta} \geq 0 \\ \xi^r, \hat{\xi}^r, \xi^i, \hat{\xi}^i \geq 0 \end{cases}
\end{aligned} \tag{38}
$$

with Wirtinger's calculus applied to RKHS described in [39], The partial derivatives of θ respect to $\mathbf{w}$, which is define at complex domain, as well as the real variables ξ^r, $\hat{\xi}^r$, ξ^i and $\hat{\xi}^i$ can be deduced

$$
\begin{cases}
\frac{\partial\Theta}{\partial\mathbf{w}^*} = \frac{1}{2}\mathbf{w} - \frac{1}{2}\sum_{k=1}^{N_r}\alpha_k\mathbf{h}_k + \frac{1}{2}\sum_{k=1}^{N_r}\hat{\alpha}_k\mathbf{h}_k + \frac{i}{2}(\sum_{k=1}^{N_r}\beta_k\mathbf{h}_k - \sum_{k=1}^{N_r}\hat{\beta}_k\mathbf{h}_k) = 0 \\
\Rightarrow \mathbf{w} = \sum_{k=1}^{N_r}(\alpha_k - \hat{\alpha}_k)\mathbf{h}_k - i\sum_{k=1}^{N_r}(\beta_k - \hat{\beta}_k)\mathbf{h}_k \\
\frac{\partial\Theta}{\partial\xi_k^r} = CR'(\xi_k^r) - \eta_k - \alpha_k = 0 \Rightarrow \eta_k = CR'(\xi_k^r) - \alpha_k \\
\frac{\partial\Theta}{\partial\hat{\xi}_k^r} = CR'(\hat{\xi}_k^r) - \hat{\eta}_k - \hat{\alpha}_k = 0 \Rightarrow \hat{\eta}_k = CR'(\hat{\xi}_k^r) - \hat{\alpha}_k \\
\frac{\partial\Theta}{\partial\xi_k^i} = CR'(\xi_k^i) - \tau_k - \beta_k = 0 \Rightarrow \tau_k = CR'(\xi_k^i) - \beta_k \\
\frac{\partial\Theta}{\partial\hat{\xi}_k^i} = CR'(\hat{\xi}_k^i) - \hat{\eta}_k - \hat{\beta}_k = 0 \Rightarrow \hat{\eta}_k = CR'(\hat{\xi}_k^i) - \hat{\beta}_k
\end{cases} \tag{39}
$$

Based on (44), we have

$$
\begin{aligned}
&< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} = \sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} + i \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} \\
&= \sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) \mathbf{h}_i \mathbf{h}_j^H + i \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) \mathbf{h}_i \mathbf{h}_j^H \\
&= \sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) \mathbf{K}_{ij}^r - \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) \mathbf{K}_{ij}^i + i (\sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) \mathbf{K}_{ij}^i + \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) \mathbf{K}_{ij}^r), \quad (40)
\end{aligned}
$$

$$
\begin{aligned}
||\mathbf{w}||_{\mathbb{H}}^2 &= \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\alpha_i - \hat{\alpha}_i) \mathbf{h}_i \mathbf{h}_j^H + \sum_{i,j=1}^{N_r} (\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_i) \mathbf{h}_i \mathbf{h}_j^H \\
&+ i (\sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j) \mathbf{h}_i \mathbf{h}_j^H - \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j) \mathbf{h}_j \mathbf{h}_i^H) \quad (41)
\end{aligned}
$$

Because $\mathbf{K}$ is Hermitian, thus $\mathbf{K}_{ij} = \mathbf{K}_{ji}^*$, if we have r_i and $r_j \in \mathbb{R}$,

$$
\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i = - \sum_{i,j}^{l} r_i r_j \mathbf{K}_{ji}^i = - \sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i, \quad (42)
$$

Therefore

$$
\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i = 0, \quad (43)
$$

Based on (48), (46) can be changed to

$$
||\mathbf{w}||_{\mathbb{H}}^2 = \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\alpha_i - \hat{\alpha}_i) \mathbf{K}_{ij}^r + \sum_{i,j=1}^{N_r} (\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_i) \mathbf{K}_{ij}^r - 2 \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j) \mathbf{K}_{ij}^i. \quad (44)
$$

Apply (44), (45), (48) and (49) to (43), the final form of Lagrange duality can be obtained

$$
\begin{aligned}
&maximize \quad \theta = -\frac{1}{2}[\sum_{i,j}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\mathbf{K}^r_{ij} + \sum_{i,j}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)\mathbf{K}^r_{ij}] \\
&- \sum_{i}^{N_r}(\alpha_i + \hat{\alpha}_i + \beta + \hat{\beta}_i)\epsilon + [\sum_{i=1}^{N_r}(\alpha_i - \hat{\alpha}_i)Re(y_i) + \sum_{i=1}^{N_r}(\beta_i - \hat{\beta}_i)Im(y_i)] \\
&+ C\sum_{i}^{N_r}(\tilde{R}(\xi_i^r) + \tilde{R}(\hat{\xi}_i^r) + \tilde{R}(\xi_i^i) + \tilde{R}(\hat{\xi}_i^i)) \\
&\begin{cases}
0 \le \alpha(\hat{\alpha}) \le C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r)) \\
0 \le \beta(\hat{\beta}) \le C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i)) \\
\xi^r(\hat{\xi}^r) \ge 0 \\
\xi^i(\hat{\xi}^i) \ge 0
\end{cases}
\end{aligned}
\tag{45}
$$

which can be divided into 2 independent regression task,

$$
\begin{aligned}
&maximize \quad \theta^r = -\frac{1}{2}\sum_{i,j}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\mathbf{K}^r_{ij} - \sum_{i=1}^{N_r}(\alpha_i + \hat{\alpha}_i)\epsilon + \sum_{i=1}^{N_r}(\alpha_i - \hat{\alpha}_i)Re(y_i) + C\sum_{i=1}^{N_r}(\tilde{R}(\xi_i^r) \\
&+ \tilde{R}(\hat{\xi}_i^r)) \\
&\begin{cases}
0 \le \alpha(\hat{\alpha}) \le C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r)) \\
\xi^r(\hat{\xi}^r) \ge 0
\end{cases}
\end{aligned}
\tag{46}
$$

$$maximize \quad \theta^i = -\frac{1}{2}\sum_{i,j}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)\mathbf{K}_{ij}^r - \sum_{i=1}^{N_r}(\beta_i + \hat{\beta}_i)\epsilon + \sum_{i=1}^{N_r}(\beta_i - \hat{\beta}_i)Im(y_i) + C\sum_{i=1}^{N_r}(\tilde{R}(\xi_i^i)$$

$$+\tilde{R}(\hat{\xi}_i^i))$$

$$\begin{cases} 0 \leq \beta(\hat{\beta}) \leq C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i)) \\ \xi^i(\hat{\xi}^i) \geq 0 \end{cases} \tag{47}$$

The alternate form can be written as

$$maximize \quad \theta^r = -\frac{1}{2}(\alpha - \hat{\alpha})^T\mathbf{K}^r(\alpha - \hat{\alpha}) + Re(\mathbf{y})^T(\alpha - \hat{\alpha}) - \epsilon(\mathbf{e}^T(\alpha + \hat{\alpha})) + C(\mathbf{e}^T(\tilde{R}(\xi^r) + \tilde{R}(\hat{\xi}^r)))$$

$$\begin{cases} 0 \leq \alpha(\hat{\alpha}) \leq C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r)) \\ \xi^r(\hat{\xi}^r) \geq 0 \end{cases} \tag{48}$$

$$maximize \quad \theta^i = -\frac{1}{2}(\beta - \hat{\beta})^T\mathbf{K}^r(\beta - \hat{\beta}) + Im(\mathbf{y})^T(\beta - \hat{\beta}) - \epsilon(\mathbf{e}^T(\beta + \hat{\beta})) + C(\mathbf{e}^T(\tilde{R}(\xi^i) + \tilde{R}(\hat{\xi}^i)))$$

$$\begin{cases} 0 \leq \beta(\hat{\beta}) \leq C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i)) \\ \xi^i(\hat{\xi}^i) \geq 0 \end{cases} \tag{49}$$

where $(\alpha - \hat{\alpha})$, $(\beta - \hat{\beta})$,$Re(\mathbf{y})$, $Im(\mathbf{y})$ denote vectors, $\mathbf{e} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{N_r}$, $\mathbf{K}^r$ denotes the matrix consist of real part of kernel components. Observe that solving (53) and (54) are equivalent to solving two independent real Support vector regression task (dual channel), only the real part of kernel matrix is required for each channel. In section VI, we will further show that from the statistic analyst of channel orthogonality (which is also named channel hardening phenomenon), the imaginary part of kernel matrix can also be omitted in stopping criteria.

Therefore, in large MIMO uplink system, our CSVR-MIMO detector can save half of the cost in kernel matrix computation.

V. Work Set Selection and Solver

(50) can be viewed as quadratic optimization problem, The traditional optimization algorithms such as Newton, Quasi Newton can not be directly applied to this problem, because the sparseness of kernel matrix $\mathbf{K}$ can not be guaranteed, so that a prohibitive storage may be required when dealing with large data set.

Decomposition method is a set of efficient algorithms that can help to conquer this difficulty. Decomposition method works iteratively, the basic idea of decomposition method is to choose a subset of variable pairs S (named work set) to optimize in each iteration step while keep the rest variable pairs N fixed. Sequential Minimal Optimization (SMO) is an extreme case of decomposition method, the work set size is 2, an analytic quadratic programming (QP) step instead of numerical QP step can be taken in each iteration.

Because (51) and (52) are symmetric, in this section we discuss real part only. By dividing the variables into work set S and fixed set N, we can divide vector α into two sub vectors, $[\alpha_S, \alpha_N]$. Thus (53) can be changed to:

$$\begin{aligned}&maximize \quad \theta^r = -\frac{1}{2}[(\alpha-\hat{\alpha})_S^T\mathbf{K}_{SS}^r(\alpha-\hat{\alpha})_S + 2(\alpha-\hat{\alpha})_N^T\mathbf{K}_{NS}^r(\alpha-\hat{\alpha})_S] + Re(\mathbf{y})_S^T(\alpha-\hat{\alpha})_S - \\ &\epsilon(\mathbf{e}^T(\alpha+\hat{\alpha})_S) - \frac{1}{2}(\alpha-\hat{\alpha})_N^T\mathbf{K}_{NN}^r(\alpha-\hat{\alpha})_N + Re(\mathbf{y})_N^T(\alpha-\hat{\alpha})_N - \epsilon(\mathbf{e}^T(\alpha+\hat{\alpha})_N) \\ &+C(\mathbf{e}^T(\tilde{R}(\xi^r) + \tilde{R}(\hat{\xi}^r))),\end{aligned} \tag{50}$$

Where $\mathbf{K}^r = \begin{bmatrix}\mathbf{K}_{SS}^r & \mathbf{K}_{SN}^r \\ \mathbf{K}_{NS}^r & \mathbf{K}_{NN}^r\end{bmatrix}$ is a permutation of $\mathbf{K}^r$, $\mathbf{K}_{SN}^r = \mathbf{K}_{NS}^r$ and $\alpha_S \in \mathbb{R}^{|S|}$ denotes the vector constitute of all the $\alpha_i \in S$. In each iteration, in (55), α_N is fixed and only the sub problem that correlated to α_S is solved i.e

$$maximize \quad \theta_S^r = -\frac{1}{2}[(\alpha-\hat{\alpha})_S^T \mathbf{K}_{SS}^r (\alpha-\hat{\alpha})_S] + [Re(\mathbf{y})_S^T - (\alpha-\hat{\alpha})_N^T \mathbf{K}_{NS}^r](\alpha-\hat{\alpha})_S - \epsilon < \mathbf{e}_S^T, (\alpha+\hat{\alpha}) >, \tag{51}$$

In decomposition method, a proper work set selection strategy is required so that speed and performance requirement can be guaranteed. One approach is to choose dual variable pairs that violate KKT conditions, so that after each iteration, the objective function can be increased according to Osuna's theorem [35], Heuristic methods are used in in order to accelerate process, in work set selection process, the algorithm first searches among the non-bound variables (that is $0 < \alpha < C\tilde{R}(\xi)$), which are more likely to violate KKT condition, then searching the whole dual variable set, the second dual variable that can maximize optimization step of the first coordinate is chosen, approximate step size is used as evaluator for sake of reducing computational cost. Lin propose another work set selection strategy based on an alternative form of KKT condition.

Another class of approaches is to choose the dual variables whose update can provide the maximum improvements to objective function. That is

$$maximize \quad \bigtriangledown \theta_S = \theta_S((\alpha + \delta_S \mathbf{e}_S), (\hat{\alpha} + \hat{\delta}_S \mathbf{e}_S)) - \theta_S(\alpha, \hat{\alpha}), \tag{52}$$

where $\delta_S = \alpha_S^{new} - \alpha_S$, the gain in (57) can be written as

$$\bigtriangledown \theta_S^r = -\frac{1}{2}[(\delta - \hat{\delta})_S^T \mathbf{K}_{SS}^r (\delta - \hat{\delta})_S + 2(\alpha - \hat{\alpha})_S^T \mathbf{K}_{SS}^r (\delta - \hat{\delta})_S] + [Re(\mathbf{y})_S^T - (\alpha - \hat{\alpha})_N^T \mathbf{K}_{NS}^r](\delta - \hat{\delta})_S$$
$$-\epsilon \mathbf{e}_S^T (\delta + \hat{\delta})_S = -\frac{1}{2}(\delta - \hat{\delta})_S^T \mathbf{K}_{SS}^r (\delta - \hat{\delta})_S + [Re(\mathbf{y})_S^T - (\alpha - \hat{\alpha})^T \mathbf{K}_S^r](\delta - \hat{\delta})_S - \epsilon \mathbf{e}_S^T (\delta + \hat{\delta})_S \tag{53}$$

In (58), we use

$$(\alpha-\hat{\alpha})_S^T\mathbf{K}_{SS}^r+(\alpha-\hat{\alpha})_N^T\mathbf{K}_{NS}^r=[(\alpha-\hat{\alpha})_S^T,(\alpha-\hat{\alpha})_N^T]\begin{bmatrix}\mathbf{K}_{SS}^r\\ \mathbf{K}_{NS}^r\end{bmatrix}=(\alpha-\hat{\alpha})^T\mathbf{K}_S^r, \tag{54}$$

where $\mathbf{K}_S^r \in \mathbb{R}^{N_r\times S}$ denotes the matrix constructed by all the columns that belong to work set S. Then we define intermediate variable vector $\Phi \in \mathbb{C}^{N_r}$, $\Phi^r = Re(\mathbf{y}) - \mathbf{K}^r(\alpha-\hat{\alpha})$ and $\Phi^i = Im(\mathbf{y}) - \mathbf{K}^r(\beta-\hat{\beta})$ Thus (58) can be rewritten as

$$\bigtriangledown\theta_S^r = -\frac{1}{2}(\delta-\hat{\delta})_S^T\mathbf{K}_{SS}^r(\delta-\hat{\delta})_S+(\Phi_S^r)^T(\delta-\hat{\delta})_S-\epsilon\mathbf{e}_S^T(\delta+\hat{\delta})_S \tag{55}$$

The offset term is omitted in Large MIMO system, therefore different from SMO type algorithms, there is no linear equation constraint as shown in (23), it is possible to update only one variable pair in each iteration. However, recent work shows more efficient work set selection strategy based on maximum gain selection approaches, that choose two pair of dual variables can reduce computational cost while maintaining the comparable performance with that with offset [36]. Here we propose sequential 1-D work set selection strategy, which can approximate the performance of optimal 2-D work set selection, while only $O(n)$ searching times required for the former one instead of $O(n^2)$ searching times.

A. Single Direction Solver

Recall KKT complementary condition

$$\begin{cases} (C\tilde{R}(\xi^r) - \alpha)\xi^r = 0 \\ (C\tilde{R}(\hat{\xi}^r) - \hat{\alpha})\hat{\xi}^r = 0 \\ \alpha(Re(y) - < \mathbf{h}, \mathbf{w} >_{\mathbb{H}} - \epsilon - \xi^r) = 0 \\ \hat{\alpha}(< \mathbf{h}, \mathbf{w} >_{\mathbb{H}} - Re(y) - \epsilon - \hat{\xi}^r) = 0 \end{cases} \tag{56}$$

it can be easily observed that $\alpha\hat{\alpha} = 0$, because $0 \leq \alpha(\hat{\alpha}) \leq C\tilde{R}(\xi^r)(C\tilde{R}(\hat{\xi}^r))$, ξ^r and $\hat{\xi}^r$ satisfy $\xi^r\hat{\xi}^r = 0$. Hence we can substitute $\lambda = \alpha - \hat{\alpha}$ and $|\lambda| = \alpha + \hat{\alpha}$, therefore the update unit is single optimization variable λ, rather than pair α and $\hat{\alpha}$. We will first introduce 1-D work set selection strategy in which one optimization variable that maximizes the gain of objective function is updated in one iteration. Reformulate by λ and $\sigma = \lambda^{new} - \lambda$ sub optimization objective function(56) and its gain (60)

$$maximize \quad \theta_S^r = -\frac{1}{2}[\lambda_S^T \mathbf{K}_{SS}^r \lambda_S] + [Re(\mathbf{y})_S^T - \lambda_N^T \mathbf{K}_{NS}^r]\lambda_S - \epsilon < \mathbf{e}_S^T, |\lambda_S| >, \tag{57}$$

$$\triangledown\theta_S^r = -\frac{1}{2}\sigma_S^T \mathbf{K}_{SS}^r \sigma_S + (\Phi_S^r)^T \sigma_S - \epsilon < \mathbf{e}_S^T, |\lambda_S^{new}| - |\lambda_S| >, \tag{58}$$

For 1-D solver, the sub optimization objective function can be written as

$$maximize \quad \theta_1^r = -\frac{1}{2}(\lambda_1^{new})^2 \mathbf{K}_{11}^r + [Re(y_1) - \sum_{j=2}^{N_r} \mathbf{K}_{1j}^r \lambda_j]\lambda_1^{new} - \epsilon(|\lambda_1^{new}|), \tag{59}$$

take the partial derivative of θ_1^r respect to λ_1^{new}, where we define $\Phi_i^r = Re(y_i) - \sum_{j=1}^{N_r} \lambda_j^r \mathbf{K}_{ij}^r$, similarly, as to dual variable λ^i , we define $\Phi_i^i = Im(y_i) - \sum_{j=1}^{N_r} \lambda_j^i \mathbf{K}_{ij}^r$. Here for sake of brevity, we use λ. Hence we have

$$
\begin{aligned}
&\frac{\partial \theta_1^r}{\partial \lambda_1^{new}} = -\lambda_1^{new}\mathbf{K}_{11}^r + Re(y_1) - \sum_{j=2}^{N_r} \lambda_j \mathbf{K}_{1j}^r - \epsilon(sgn(\lambda_1^{new})) = \\
&-\lambda_1^{new}\mathbf{K}_{11}^r + \Phi_1^r + \lambda_1 \mathbf{K}_{11}^r - \epsilon(sgn(\lambda_1^{new})) \\
&\Rightarrow \lambda_1^{new} = \lambda_1 + \frac{\Phi_1^r - \epsilon(sgn(\lambda_1^{new}))}{\mathbf{K}_{11}^r},
\end{aligned} \tag{60}
$$

The update of α_1 or $\hat{\alpha}_1$ is completed by clipping

$$
\lambda_1^{new \quad clipped} = [\lambda_1^{new}]_{-CR'(\xi)}^{CR'(\xi)} \tag{61}
$$

where $[]_a^b$ denotes clipping function

$$
[x]_a^b = \begin{cases} a \quad if \quad x \leq a \\ x \quad if \quad a < x < b \\ b \quad if \quad x \geq b \end{cases} \tag{62}
$$

Define $\sigma = \lambda^{new \quad clipped} - \lambda$, based on (63), The gain of objective function respect to ith optimization variable is

$$
\begin{aligned}
&\bigtriangledown \theta_i^r = \theta^r(\lambda_1 + \sigma_1) - \theta^r(\lambda_1) \\
&= -\frac{1}{2}\sigma_i^2 \mathbf{K}_{ii}^r + \Phi_i^r \sigma_i - \epsilon(|\lambda_i^{new \quad cipped}| - |\lambda_i|) \\
&= \sigma_i[-\frac{1}{2}\sigma_i \mathbf{K}_{ii}^r + \Phi_i^r] - \epsilon(|\lambda_i^{new \quad clipped}| - |\lambda_i|),
\end{aligned} \tag{63}
$$

In 1-D searching procedure, the optimization variable which has the maximum gain of sub optimization objective function is updated as 1 in (64), that is

$$1 = arg_{(i=1,\ldots,N_r)} \max \bigtriangledown \Theta_i^r, \tag{64}$$

B. Double Direction Solver

Although omission of offset in the CSVR-MIMO detector makes 1-D solver possible, however recent work in machine learning field shows training SVM without offset by 2-D solver with special work set selection strategies has more rapid training speed while the comparable performance is retained. The 2-D solver uses the same principle as 1-D solver, the work set size is 2, that is $\lambda_S = \lambda_1 \mathbf{e}_1 + \lambda_2 \mathbf{e}_2$. Based on (63), the sub objective function can be written as

$$\begin{aligned}
&maximize \quad \theta_{1,2}^r = -\frac{1}{2}[(\lambda_1^{new})^2 \mathbf{K}_{11}^r + (\lambda_2^{new})^2 \mathbf{K}_{22}^r + 2\lambda_1^{new}\lambda_2^{new}\mathbf{K}_{12}^r] - \\
&\lambda_1^{new} \sum_{j \neq 1,2}^{N_r} \lambda_j \mathbf{K}_{1j}^r - \lambda_2^{new} \sum_{j \neq 1,2}^{N_r} \lambda_j \mathbf{K}_{2j}^r + Re(y_1)\lambda_1^{new} + Re(y_2)\lambda_2^2 \\
&-\epsilon(|\lambda_1^{new}| + |\lambda_2^{new}|),
\end{aligned} \tag{65}$$

Based on (70), the partial derivatives of $\theta_{1,2}^r$ with respect to λ_1^{new} and λ_2^{new} are

$$\begin{aligned}
&\frac{\partial \theta_{1,2}^r}{\partial \lambda_1^{new}} = -\lambda_1^{new}\mathbf{K}_{11}^r - \lambda_2^{new}\mathbf{K}_{12}^r - \sum_{j \neq 1,2}^{N_r} \lambda_j \mathbf{K}_{1j}^r + Re(y_1) - \epsilon sgn(\lambda_1^{new}) = \\
&-\lambda_1^{new}\mathbf{K}_{11}^r - \lambda_2^{new}\mathbf{K}_{12}^r + \Phi_1^r + \lambda_1\mathbf{K}_{11}^r + \lambda_2\mathbf{K}_{12}^r - \epsilon sgn(\lambda_1^{new}) = 0
\end{aligned} \tag{66}$$

$$\begin{aligned}
&\frac{\partial \theta_{1,2}^r}{\partial \lambda_2^{new}} = -\lambda_2^{new}\mathbf{K}_{22}^r - \lambda_1^{new}\mathbf{K}_{12}^r - \sum_{j \neq 1,2}^{N_r} \lambda_j \mathbf{K}_{2j}^r + Re(y_2) - \epsilon sgn(\lambda_2^{new}) = \\
&-\lambda_2^{new}\mathbf{K}_{22}^r - \lambda_1^{new}\mathbf{K}_{12}^r + \Phi_2^r + \lambda_1\mathbf{K}_{12}^r + \lambda_2\mathbf{K}_{22}^r - \epsilon sgn(\lambda_2^{new}) = 0
\end{aligned} \tag{67}$$

where $\frac{\partial |x|}{\partial x} = sgn(x)$ denotes the sign of x. Based on (71) and (72) we have

$$(\lambda_1^{new} - \lambda_1)\mathbf{K}_{11}^r = \Phi_1^r - \epsilon sgn(\lambda_1^{new}) - (\lambda_2^{new} - \lambda_2)\mathbf{K}_{12}^r \tag{68}$$

$$(\lambda_2^{new} - \lambda_2)\mathbf{K}_{22}^r = \Phi_2^r - \epsilon sgn(\lambda_2^{new}) - (\lambda_1^{new} - \lambda_1)\mathbf{K}_{12}^r \tag{69}$$

hence based on (73) and (74), the update formula of λ_1^{new} and λ_2^{new} are

$$\lambda_1^{new} = \lambda_1 + \frac{\Phi_1^r\mathbf{K}_{22}^r - \Phi_2^r\mathbf{K}_{12}^r - \epsilon[sgn(\lambda_1^{new})\mathbf{K}_{22}^r - sgn(\lambda_2^{new})\mathbf{K}_{12}^r]}{\mathbf{K}_{11}^r\mathbf{K}_{22}^r - (\mathbf{K}_{12}^r)^2} \tag{70}$$

$$\lambda_2^{new} = \lambda_2 + \frac{\Phi_2^r\mathbf{K}_{11}^r - \Phi_1^r\mathbf{K}_{12}^r - \epsilon[sgn(\lambda_2^{new})\mathbf{K}_{11}^r - sgn(\lambda_1^{new})\mathbf{K}_{12}^r]}{\mathbf{K}_{11}^r\mathbf{K}_{22}^r - (\mathbf{K}_{12}^r)^2} \tag{71}$$

Then the updated optimization variables are clipped by constraint

$$\lambda^{new \quad clipped} = [\lambda^{new}]_{-CR'(\xi)}^{CR'(\xi)}, \tag{72}$$

It is obviously the dual variables in 2-D solver have the same update rule as that of 1-D solver. Based on (60), assume the ith and jth dual variable pair are chosen, the gain of 2-D solver objective function can be written as

$$\begin{aligned}\triangledown\theta_{ij}^r = &-\frac{1}{2}[\sigma_i^2\mathbf{K}_{ii}^r + \sigma_j^2\mathbf{K}_{jj}^r + 2\sigma_i\sigma_j\mathbf{K}_{ij}^r] + \Phi_i^r\sigma_i + \Phi_j^r\sigma_j \\ &-\epsilon(|\lambda_i^{new \quad clipped}| - |\lambda_i| + |\lambda_j^{new \quad clipped}| - |\lambda_j|),\end{aligned} \tag{73}$$

recall the gain of objective function of 1-D solver in (68), we obtain

$$\bigtriangledown\theta_{ij}^r = \bigtriangledown\theta_i^r + \bigtriangledown\theta_j^r - \sigma_i\sigma_j\mathbf{K}_{ij}^r, \tag{74}$$

where $\bigtriangledown\Theta_i^r$, $\bigtriangledown\Theta_j^r$ denote gains of 1-D solver with ith and jth dual variable pairs are chosen.

C. Approximation to Optimal Double Direction Solver based on Single Direction Solver

From (79), it is obviously that the gain of 2-D solver is a summation of the gain of 2 independent 1-D solver and a correlation term $\sigma_i\sigma_j\mathbf{K}_{ij}^r$.

Obviously optimal 2-D coordinate combination (i, j) can be determined by comparing the gains of all the possibilities exhaustively, which requires $O(n^2)$ times of searching. Based on (79), we can approximate optimal 2-D solution by 1-D search approach, we will prove in large MIMO scenario, when N_t is sufficient large, with channel hardening become effective, this approximation is very efficient. Here we propose two kinds of 1-D approximate searching strategy:

1) 1-D searching without damping: do one round 1-D searching and calculate all the 1-D gain based on (68), then choose the coordinates with first and second largest 1-D gain as the candidates, then update the two candidates by 2-D solver as shown in (75) and (76)

2) 1-D searching with damping: do two rounds 1-D searchings, in the first round find optimization variable i that can maximize 1-D gain, then in the second round, find jth optimization variable with the value of ith coordinate updated.

From (79), it can be easily interpreted the efficient of 1-D approximation approach is majorly determined by the approximation ratio $\frac{\sigma_i\sigma_j\mathbf{K}_{ij}^r}{\bigtriangledown\theta_i^r+\bigtriangledown\theta_j^r}$, hence we provide theoretical analyse from the view of channel hardening phenomenon. Prior the theoretical analyse, we first investigate some mathematical properties of channel hardening (to be completed).

For 1-D solver the gradient of (65) with respect to λ can be written as

$$\lambda_1^{new} = \lambda_1 + \frac{\Phi_1 - sgn(\lambda_1^{new})\epsilon}{\mathbf{K}_{11}^r}, \tag{75}$$

In the update process the $sgn(\lambda^{new})$ is unknown at current step, therefore, we need to consider both the case $sgn(\lambda^{new}) = -1$ or 1 and choose the one with the larger objective function gain $\bigtriangledown\theta_S^r$.

VI. Stopping Criteria

As we have explained in section II-C, the upper bound of Lagrangian dual objective function is determined by primal objective function, further more the optimal of primal and dual objective function is found if and only if the equality holds, that is

$$\theta(\lambda^r, \lambda^i) = f(\mathbf{w}, \xi) \tag{76}$$

$$\frac{1}{2}||\mathbf{w}||_{\mathbb{H}}^2 + C\sum_{i=1}^{N_r}[R(\xi_i^r) + R(\hat{\xi}_i^r) + R(\xi_i^i) + R(\hat{\xi}_i^i)], \tag{77}$$

(50) can be rewritten as follow by substituting $\lambda^r = \alpha - \hat{\alpha}$, $|\lambda^r| = \alpha + \hat{\alpha}$ and $\lambda^i = \beta - \hat{\beta}$, $|\lambda^i| = \beta + \hat{\beta}$

$$\begin{aligned}\theta(\lambda^r, \lambda^i) = &-\frac{1}{2} < (\lambda^r)^T, \mathbf{K}^r\lambda^r > -\frac{1}{2} < (\lambda^i)^T, \mathbf{K}^r\lambda^i > + < Re(\mathbf{y})^T, \lambda^r > + < Im(\mathbf{y})^T, \lambda^i > \\ &-\epsilon < \mathbf{e}^T, (|\lambda^r| + |\lambda^i|) > +C\sum_{i=1}^{N_r}[\tilde{R}(\xi_i^r) + \tilde{R}(\hat{\xi}_i^r) + \tilde{R}(\xi_i^i) + \tilde{R}(\hat{\xi}_i^i)],\end{aligned} \tag{78}$$

Similarly, (46) can be formulated as

$$||\mathbf{W}||_{\mathbb{H}}^2 =< (\lambda^r)^T, \mathbf{K}^r \lambda^r > + < (\lambda^i)^T, \mathbf{K}^r \lambda^i > -2 < \lambda^r, \mathbf{K}^i \lambda^i >, \tag{79}$$

hence, duality gap can be formulated as

$$G(\lambda^r, \lambda^i) = f(\mathbf{w}, \xi) - \theta(\lambda^r, \lambda^i) =< (\lambda^r)^T, \mathbf{K}^r \lambda^r > + < (\lambda^i)^T, \mathbf{K}^r \lambda^i > - < Re(\mathbf{y})^T, \lambda^r > - < Im(\mathbf{y})^T, \lambda^i >$$
$$-\epsilon < \mathbf{e}^T, (|\lambda^r| + |\lambda^i|) > +C \sum_{i=1}^{N_r} [\xi_i^r R^{'}(\xi_i^r) + \hat{\xi}_i^r R^{'}(\hat{\xi}_i^r) + \xi_i^i R^{'}(\xi_i^i) + \hat{\xi}_i^i R^{'}(\hat{\xi}_i^i)] - 2 < \lambda^r, \mathbf{K}^i \lambda^i > . \tag{80}$$

As we explained in section **??**, the choice of risk function is determined by distribution of noise, as to Gaussian noise, the risk function is

$$R(\xi) = \frac{1}{2} \xi^2, \tag{81}$$

hence

$$\tilde{R}(\xi) = R(\xi) - \xi R'(\xi) = -\frac{1}{2} \xi^2, \tag{82}$$

In ϵ-SVR, the objective to employ slack variables ξ is to deal with the outliers that outside ϵ tube to compensate the influence from noise. Therefore

$$\xi_i^r = Re(\mathbf{y}_i) - Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - \epsilon \tag{83}$$

$$\hat{\xi}_i^r = Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - Re(\mathbf{y}_i) - \epsilon \tag{84}$$

Because $\xi^r \hat{\xi}^r = 0$ (estimation can only exceed ϵ tube in one direction), therefore there is only one of ξ and $\hat{\xi}$ need to be considered, thus

$$\xi_i^r(\hat{\xi}_i^r) = \max(0, |Re(\mathbf{y}_i) - Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})| - \epsilon) \tag{85}$$

$$\xi_i^i(\hat{\xi}_i^i) = \max(0, |Im(\mathbf{y}_i) - Im(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})| - \epsilon) \tag{86}$$

$$\xi \hat{\xi} = 0, \tag{87}$$

Therefore the risk function can be rewritten as

$$R(\xi_i^r) + R(\hat{\xi}_i^r) = \frac{1}{2}(|Re(\mathbf{y}_i) - Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})|)_\epsilon^2 \tag{88}$$

$$R(\xi_i^i) + R(\hat{\xi}_i^i) = \frac{1}{2}(|Im(\mathbf{y}_i) - Im(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})|)_\epsilon^2 \tag{89}$$

where $()_\epsilon$ denotes ϵ insensitive function as we mention in section **??**. Based on (45), we have

$$Re(\mathbf{y}_i) - Re(< \mathbf{h}_i, \mathbf{W} >_{\mathbb{H}}) = Re(\mathbf{y}_i) - \sum_{j=1}^{N_r} \lambda_j^r \mathbf{K}_{ij}^r + \sum_{j=1}^{N_r} \lambda_j^i \mathbf{K}_{ij}^i \tag{90}$$

$$Im(\mathbf{y}_i) - Im(< \mathbf{h}_i, \mathbf{W} >_{\mathbb{H}}) = Im(\mathbf{y}_i) - \sum_{j=1}^{N_r} \lambda_j^i \mathbf{K}_{ij}^r - \sum_{j=1}^{N_r} \lambda_j^r \mathbf{K}_{ij}^i \tag{91}$$

we define two intermediate variables Φ and Ψ

$$\Phi^r = Re(\mathbf{y}) - \mathbf{K}^r \lambda^r; \Phi^i = Im(\mathbf{y}) - \mathbf{K}^r \lambda^i \tag{92}$$

$$\Psi^r = \mathbf{K}^i \lambda^i; \Psi^i = -\mathbf{K}^i \lambda^r \tag{93}$$

Therefore based on (94)-(98), duality gap in (85) can be rewritten as

$$G(\lambda^r, \lambda^i) =< (\lambda^r)^T, \mathbf{K}^r \lambda^r > + < (\lambda^i)^T, \mathbf{K}^r \lambda^i > - < Re(\mathbf{y})^T, \lambda^r > - < Im(\mathbf{y})^T, \lambda^i >$$
$$+\epsilon < \mathbf{e}^T, (|\lambda^r| + |\lambda^i|) > +C \sum_{i=1}^{N_r} [(|\Phi_i^r + \Psi_i^r|)_\epsilon^2 + (|\Phi_i^i + \Psi_i^i|)_\epsilon^2] - 2 < \lambda^r, \mathbf{K}^i \lambda^i > . \tag{94}$$

Based on objective function in (83), (99) can be rewritten as

$$G = (\lambda_r, \lambda_i) =< Re(\mathbf{y})^T, \lambda^r > + < Im(\mathbf{y})^T, \lambda^i > -\epsilon < \mathbf{e}^T, (|\lambda^r| + |\lambda^i|) > -2\theta(\lambda_i, \lambda_j)$$
$$-2 < \lambda^r, \mathbf{K}^i \lambda^i > . \tag{95}$$

The duality gap between primal problem and dual problem is used to evaluate how close a solution is to global minimum. In our scenario, duality gap is employed as stopping criteria. Therefore to make stopping criteria more effective to monitor if algorithm convergent, we monitor the ratio by a value of tolerance (usually this tolerance is set to 10^{-3}).

$$\frac{G}{G + \theta} \tag{96}$$

A. *Update* Φ, Ψ *and* G

In each iteration Φ, Ψ and G are updated partially based on 2 updated optimization variables. Here we give the pseudo code to update Φ, Ψ and G.

Based on the definition of Φ and Ψ in (97) and (98), we have the following procedure to update Φ and Ψ in real channel and imaginary channel, assume the optimization coordinate updated in each channel are 1 and 2.

procedure 1. UPDATE Φ^r AND Ψ^i IN REAL CHANNEL
  for $i = 1 : N_r$ **do**
    $\Phi_i^r = \Phi_i^r - \sigma_1^r \mathbf{K}_{i1}^r - \sigma_2^r \mathbf{K}_{i2}^r$
    $\Psi_i^i = \Psi_i^i - \sigma_1^r \mathbf{K}_{i1}^i - \sigma_2^r \mathbf{K}_{i2}^i$
  end for
end procedure

procedure 2. UPDATE Φ^i AND Ψ^r IN IMAGINARY CHANNEL
  for $i = 1 : N_r$ **do**
    $\Phi_i^i = \Phi_i^i - \sigma_1^i \mathbf{K}_{i1}^r - \sigma_2^i \mathbf{K}_{i2}^r$
    $\Psi_i^r = \Psi_i^r + \sigma_1^i \mathbf{K}_{i1}^i + \sigma_2^i \mathbf{K}_{i2}^i$
  end for
end procedure

Then the risk function term in (99) is updated as following

procedure 3. UPDATE RISK FUNCTION IN REAL CHANNEL(χ^r)
  $\chi^r = 0$ ▷ initial risk term
  for $i = 1 : N_r$ **do**
    if $|\Phi_i^r + \Psi_i^r| > \epsilon$ **then**
      $\chi^r += (|\Phi_i^r + \Psi_i^r| - \epsilon)^2$
    end if
  end for
end procedure

procedure 4. UPDATE RISK FUNCTION IN IMAGINARY CHANNEL(χ^i)
  $\chi^i = 0$ ▷ initial risk term
  for $i = 1 : N_r$ **do**
    if $|\Phi_i^i + \Psi_i^i| > \epsilon$ **then**
      $\chi^i += (|\Phi_i^i + \Psi_i^i| - \epsilon)^2$
    end if
  end for
end procedure

The pseudo code to update duality gap G based on (99) is shown as follow assume the coordinate updated in real channel is i and j, in imaginary channel is m and f.

procedure 5. UPDATE G

$G+ = Re(\mathbf{y}_1)\sigma_i^r + Re(\mathbf{y}_2)\sigma_j^r$

$G+ = Im(\mathbf{y}_1)\sigma_m^i + Re(\mathbf{y}_2)\sigma_f^i$

$G- = \epsilon(|\lambda_i^r + \sigma_i^r| - |\lambda_i^r| + |\lambda_j^r + \sigma_j^r| - |\lambda_j^r|)$

$G- = \epsilon(|\lambda_m^i + \sigma_m^i| - |\lambda_m^i| + |\lambda_f^i + \sigma_f^i| - |\lambda_f^i|)$

$G- = 2(\bigtriangledown\theta_{i,j,m,f}(\sigma_i^r, \sigma_j^r, \sigma_m^i, \sigma_f^i))$ ▷ Update sub objective function based on (79)

$G+ = C(\chi^r + \chi^i)^{new} - (\chi^r + \chi^i)$ ▷ Update risk function term based on **Procedure 3** and **Procedure 4**

$G- = \sigma_i^r\sigma_m^i\mathbf{K}_{im}^i + \sigma_i^r\sigma_f^i\mathbf{K}_{if}^i + \sigma_j^r\sigma_m^i\mathbf{K}_{jm}^i + \sigma_j^r\sigma_f^i\mathbf{K}_{jf}^i + \sigma_m^i\sum_{k=1}^{N_r}\lambda_k^r\mathbf{K}_{km}^i + \sigma_f^i\sum_{k=1}^{N_r}\lambda_k^r\mathbf{K}_{kf}^i - \sigma_i^r\sum_{k=1}^{N_r}\lambda_k^i\mathbf{K}_{ki}^i - \sigma_j^r\sum_{k=1}^{N_r}\lambda_k^i\mathbf{K}_{kj}^i$ ▷ Update $< \lambda^r, \mathbf{K}^i\lambda^i >$

end procedure

Pseudo code for sequential single searching 2-D solver is shown as following

procedure 6. SEQUENTIAL SINGLE SEARCHING 2-D SOLVER WITHOUT DAMPING

Step 1. Search for two optimization variables based on single direction solver

for $i = 1 : N_r$ **do**

calculate $\bigtriangledown\theta_i^r(\bigtriangledown\theta_i^i)$ ▷ Based on single direction solver V-A

end for

choose the dual variable with first and the second largest gain of sub objective function, denoted as 1st and 2nd

Step 2. Update 1st and 2nd optimization variables based on double direction solver

update $\lambda_{1st}^r(\lambda_{1st}^i)$ and $\lambda_{2nd}^r(\lambda_{2nd}^i)$ ▷ Based on double direction solver V-B

update $\Phi^r(\Phi^i)$ and $\Psi^r(\Psi^i)$ by **Procedure 1** and **Procedure 2**

end procedure

The following is the pseudo code of complex support vector detector (CSVD) is shown in Appendix A

procedure 7. SEQUENTIAL SINGLE SEARCHING 2-D SOLVER WITH DAMPING
Step 1. Search for two optimization variables based on single direction solver
for $i = 1 : N_r$ **do** ▷ First round searching
calculate $\bigtriangledown\theta_i^r(\bigtriangledown\theta_i^i)$ ▷ Based on single direction solver V-A
end for
choose the optimization variable with the largest gain of objective function as $1st_1$
update $\Phi^r(\Phi^i)$ and $\Psi^r(\Psi^i)$ with respect to $1st_1$
for $i = 1 : N_r$ **do** ▷ Second round searching
calculate $\bigtriangledown\theta_i^r(\bigtriangledown\theta_i^i)$ ▷ Based on single direction solver V-A
end for
choose the optimization variable with the largest gain of objective function as $1st_2$
Step 2. Update $1st_1$ and $1st_2$ optimization variables based on double direction solver
update $\lambda^r_{1st_1}(\lambda^i_{1st_1})$ and $\lambda^r_{1st_2}(\lambda^i_{1st_2})$ ▷ Based on double direction solver V-B
update $\Phi^r(\Phi^i)$ and $\Psi^r(\Psi^i)$ by **Procedure 1** and **Procedure 2**
end procedure

VII. COMPUTER SIMULATIONS

Computer simulation is launched to test the detection and run time performance of proposed dual channel complex support vector detection algorithm. For sake of brevity, the real case is tested first, all the experiments are made by C, compiled by gcc version 4.8.3 on 64 bit Fedora (release 19) Linux system. The experiment platform is a desktop computer with I5-4th generation CPU with quad processing cores, 3.2 GHz clock rate, 8 GB RAM.

For sake of brevity, we consider a real uncoded spatial multiplex large MIMO system to simulate one channel of the proposed dual channel complex support vector detection algorithm. with N_r received antennas and N_t transmitted antennas. The propagation channel matrix is constructed by channel gain components that are identically independent distributed (i.i.d) Gaussian random variables with zero mean and unit variance. transmitted symbols are mutually independent modulated by M PAM with normalized average energy $\frac{1}{N_t}$, transmitted over flat fading channel, the sample of noise is AWGN with zero mean and variance $\frac{1}{10^{SNR/10}}$, where SNR

denotes the signal to noise ratio. We make experiment to low loading factor system 100×40 and full loading factor 100×100, with at least $1e^5$ channel realizations and at least 500 symbol errors accumulated. Fig.3 shows the symbol error rate (SER) performance, Table.I shows the average iteration time of real SVD for different SNR.

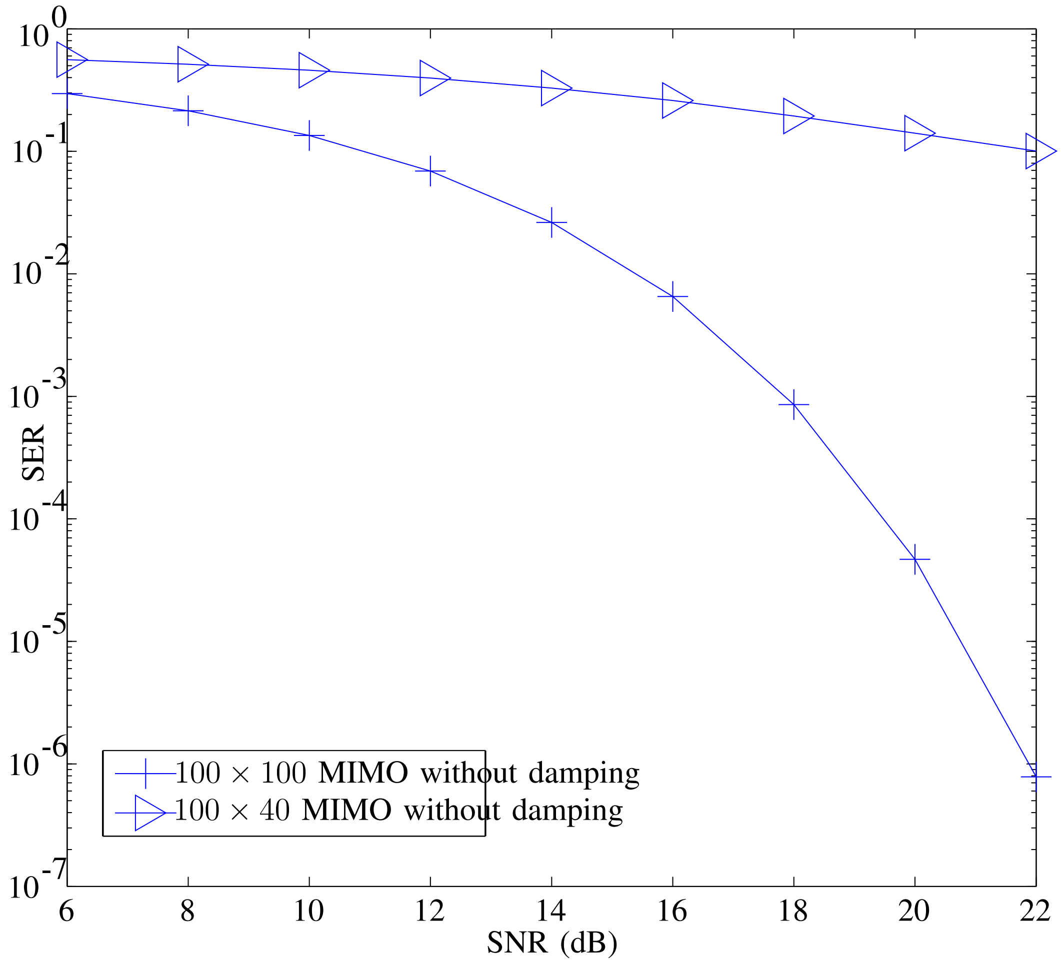

Fig. 3. SER performance of 100×100 and 100×40 MIMO system

TABLE I
AVERAGE ITERATION TIME OF REAL SUPPORT VECTOR DETECTOR

Array Size	**SNR**								
	6	8	10	12	14	16	18	20	22
100×40	682	681	681	681	680	679	678	677	680
100×100	1925	1916	1903	1885	1862	1827	1782	1723	1654

APPENDIX A

PSEUDO CODE OF CSVD

Algorithm 1 Dual Channel Complex Support Vector Detection Algorithm

procedure CSVD($\mathbf{y}$,$\mathbf{H}$)
  Step 1. Initialization
  $\mathbf{K} = \mathbf{H}\mathbf{H}^H$ ▷ kernel matrix
  $\chi^r = 0$, $\chi^i = 0$ ▷ risk function
  for $i = 1 : N_r$ **do** ▷ initialize λ^r, λ^i, Φ^r, Φ^i, Ψ^r, Ψ^i and duality gap G
    $\lambda_i^r = 0, \lambda_i^i = 0$
    $\Phi_i^r = Re(y_i), \Phi_i^i = Im(y_i)$
    $\Psi_i^r = 0, \Psi_i^i = 0$
    if $|\Phi_i^r| > \epsilon$ **then**
      $\chi^r += (|\Phi_i^r| - \epsilon)^2$
    end if
    if $|\Phi_i^i| > \epsilon$ **then**
      $\chi^i += (|\Phi_i^i| - \epsilon)^2$
    end if
  end for
  $G = C(\chi^r + \chi^i)$ ▷ initialize duality gap
  $\theta = -0.5G$ ▷ initialize objective function
  Step 2. if $G > tol$, go to step 3, else go to Step 5
  Step 3.
  Sequentia single searching 2-D solver with or without damping ▷ find two optimization variables to be updated
  Step 4. **Procedure 5** update G
  Step 5.
  $\tilde{x} = (\lambda^r + i\lambda^i)^T\mathbf{H}$ ▷ reconstruct $\mathbf{x}$
  $\mathbf{x} = \mathbb{Q}(\tilde{x})$ ▷ $\mathbb{Q}(\cdot)$ denotes quantization operation based on symbol constellation
  go back to Step 2
  Step 6. **Return** $\mathbf{x}$
end procedure

REFERENCES

[1] "IEEE Standard for Information technology– Telecommunications and information exchange between systemslocal and metropolitan area networks– Specific requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications–Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz." *IEEE Std 802.11ac-2013 (Amendment to IEEE Std 802.11-2012, as amended by IEEE Std 802.11ae-2012, IEEE Std 802.11aa-2012, and IEEE Std 802.11ad-2012)*, pp. 1–425, Dec 2013.

[2] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband*. Academic press, 2010.

[3] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *Signal Processing Magazine, IEEE*, vol. 30, no. 1, pp. 40–60, 2013.

[4] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems," *Communications Magazine, IEEE*, vol. 52, no. 2, pp. 186–195, 2014.

[5] P. W. Wolniansky, G. J. Foschini, G. Golden, R. Valenzuela *et al.*, "V-BLAST: An architecture for realizing very high data rates over the rich-scattering wireless channel," in *Signals, Systems, and Electronics, 1998. ISSSE 98. 1998 URSI International Symposium on*. IEEE, 1998, pp. 295–300.

[6] G. J. Foschini, G. D. Golden, R. Valenzuela, P. W. Wolniansky *et al.*, "Simplified processing for high spectral efficiency wireless communication employing multi-element arrays," *Selected Areas in Communications, IEEE Journal on*, vol. 17, no. 11, pp. 1841–1852, 1999.

[7] J. Benesty, Y. Huang, and J. Chen, "A fast recursive algorithm for optimum sequential signal detection in a blast system," *Signal Processing, IEEE Transactions on*, vol. 51, no. 7, pp. 1722–1730, 2003.

[8] M. O. Damen, H. El Gamal, and G. Caire, "On maximum-likelihood detection and the search for the closest lattice point," *Information Theory, IEEE Transactions on*, vol. 49, no. 10, pp. 2389–2402, 2003.

[9] J. Jaldén and B. Otterste, "On the complexity of sphere decoding in digital communications," *Signal Processing, IEEE Transactions on*, vol. 53, no. 4, pp. 1474–1484, 2005.

[10] L. G. Barbero and J. S. Thompson, "Fixing the complexity of the sphere decoder for MIMO detection," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 6, pp. 2131–2142, 2008.

[11] Z. Luo, M. Zhao, S. Liu, and Y. Liu, "Generalized parallel interference cancellation with near-optimal detection performance," *Signal Processing, IEEE Transactions on*, vol. 56, no. 1, pp. 304–312, 2008.

[12] D. Radji and H. Leib, "Interference cancellation based detection for v-blast with diversity maximizing channel partition," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 6, pp. 1000–1015, 2009.

[13] K. V. Vardhan, S. K. Mohammed, A. Chockalingam, and B. S. Rajan, "A low-complexity detector for large MIMO systems and multicarrier CDMA systems," *Selected Areas in Communications, IEEE Journal on*, vol. 26, no. 3, pp. 473–485, 2008.

[14] P. Li and R. D. Murch, "Multiple output selection-LAS algorithm in large MIMO systems," *Communications Letters, IEEE*, vol. 14, no. 5, pp. 399–401, 2010.

[15] N. Srinidhi, T. Datta, A. Chockalingam, and B. S. Rajan, "Layered tabu search algorithm for large-MIMO detection and a lower bound on ML performance," *Communications, IEEE Transactions on*, vol. 59, no. 11, pp. 2955–2963, 2011.

[16] T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Random-restart reactive tabu search algorithm for detection in

large-mimo systems," *Communications Letters, IEEE*, vol. 14, no. 12, pp. 1107–1109, 2010.

[17] P. Som, T. Datta, N. Srinidhi, A. Chockalingam, and B. S. Rajan, "Low-complexity detection in large-dimension MIMO-ISI channels using graphical models," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 5, no. 8, pp. 1497–1511, 2011.

[18] P. Som, T. Datta, A. Chockalingam, and B. S. Rajan, "Improved large-MIMO detection based on damped belief propagation," in *Information Theory Workshop (ITW), 2010 IEEE*. IEEE, 2010, pp. 1–5.

[19] T. L. Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing (CHEMP) receiver in large-scale MIMO systems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 5, pp. 847–860, 2014.

[20] J. Goldberger and A. Leshem, "MIMO detection for high-order QAM based on a Gaussian tree approximation," *Information Theory, IEEE Transactions on*, vol. 57, no. 8, pp. 4973–4982, 2011.

[21] S. Mohammed, A. Chockalingam, and B. S. Rajan, "Low-complexity near-map decoding of large non-orthogonal stbcs using pda," in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*. IEEE, 2009, pp. 1998–2002.

[22] T. Datta, N. A. Kumar, A. Chockalingam, and B. S. Rajan, "A novel monte-carlo-sampling-based receiver for large-scale uplink multiuser MIMO systems," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 7, pp. 3019–3038, 2013.

[23] Q. Zhou and X. Ma, "Element-based lattice reduction algorithms for large MIMO detection," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 2, pp. 274–286, 2013.

[24] V. Vapnik, "Pattern recognition using generalized portrait method," *Automation and remote control*, vol. 24, pp. 774–780, 1963.

[25] V. Vapnik and A. Chervonenkis, "A note on one class of perceptrons," *Automation and remote control*, vol. 25, no. 1, 1964.

[26] B. Schölkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[27] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[28] I. Guyon, B. Boser, and V. Vapnik, "Automatic capacity tuning of very large vc-dimension classifiers," *Advances in neural information processing systems*, pp. 147–147, 1993.

[29] V. Vapnik, *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[31] B. Schölkopf, C. Burges, and V. Vapnik, "Incorporating invariances in support vector learning machines," in *Artificial*

Neural NetworksICANN 96. Springer, 1996, pp. 47–52.

[32] V. Vapnik, S. E. Golowich, and A. Smola, “Support vector method for function approximation, regression estimation, and signal processing,” in *Advances in Neural Information Processing Systems 9*. Citeseer, 1996.

[33] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[34] J. Platt *et al.*, “Fast training of support vector machines using sequential minimal optimization,” *Advances in kernel methodssupport vector learning*, vol. 3, 1999.

[35] E. Osuna, R. Freund, and F. Girosi, “An improved training algorithm for support vector machines,” in *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*. IEEE, 1997, pp. 276–285.

[36] I. Steinwart, D. Hush, and C. Scovel, “Training svms without offset,” *The Journal of Machine Learning Research*, vol. 12, pp. 141–202, 2011.

[37] P. Bouboulis, S. Theodoridis, C. Mavroforakis, and L. Evaggelatou-Dalla, “Complex support vector machines for regression and quaternary classification.”

[38] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[39] P. Bouboulis and S. Theodoridis, “Extension of Wirtinger’s calculus to reproducing kernel Hilbert spaces and the complex kernel LMS,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 3, pp. 964–978, 2011.