Report: Complex Support Vector Detector for Large MIMO System

Tianpei Chen

Department of Electrical and Computer Engineering

McGill University

October 18, 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 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] and 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 spectrum efficiency) and diversity gain (better reliability of communication link) are 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 to both 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 ends. The uplink detector is one of the key components in a large MIMO system. 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, complexities of the algorithms become 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 the 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 the complexity of SD still has a lower bound 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 values 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 metaheuristic based 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 performances compared to LAS detectors because of the efficient local minima exit strategy (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]) . Moreover, element based Lattice Reduction (LR) aided algorithms were also proposed to large MIMO systems. [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] were proposed which solve nonlinear learning tasks by mapping input data samples 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 form, which can be solved by Quadratic Programming (QP). Decomposition methods were proposed to solve this QP problem by decomposing it into sub QP problems and solving them iteratively [34]. Therefore, the computational intensive numerical methods can be avoided. Decomposition is performed by sub set selection solver, which refers to a set of algorithms that separate the optimization variables (Lagrange multipliers) into two sets S and N, S is the work set and N contains the remaining optimization variables. In each iteration, only the optimization variables in the work set is updated while keeping other optimization variables fixed. The Sequential Minimal Optimization (SMO) algorithm [34] is an extreme case of the decomposition solvers. An important issue of the sub set selection solvers is the selection of the work set. One strategy is to choose Karush-Kuhn-Tucker (KKT) condition violators, ensuring the final converge [35]. The SMO algorithm restricts the size of the work set to two. A method to train SVM without offset was proposed In [36], 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 different 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 the better performance as well as reduced complexity comparing to simply split learning task into two real case by real kernels. Based on this work, we derive 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 cost saving can be achieved. 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, Two types of combined single optimization variable selection strategy are proposed based on the work in [36]. The proposed combined single optimization variable selection strategy can approximate optimal double optimization variable selection strategy. The former one can achieve as few as iteration time while enjoy significant speed gain in each iteration.

II. Brief Introduction to ϵ-Support Vector Regression

A. Regression Model

Suppose we are given a 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, $\mathbf{w}, \Phi(\mathbf{x}_i) \in \mathbb{R}^\Omega$, $\Omega \in \mathbb{R}$ denotes the dimension of mapped feature space (For linear model, $\mathbf{x}_i = \Phi(\mathbf{x}_i)$, $\Omega = V$). The regression model (either linear or non-linear) is given by

$$y_i = \mathbf{w}^T \Phi(\mathbf{x}_i) + b \quad i = 1, 2, \ldots, L \tag{1}$$

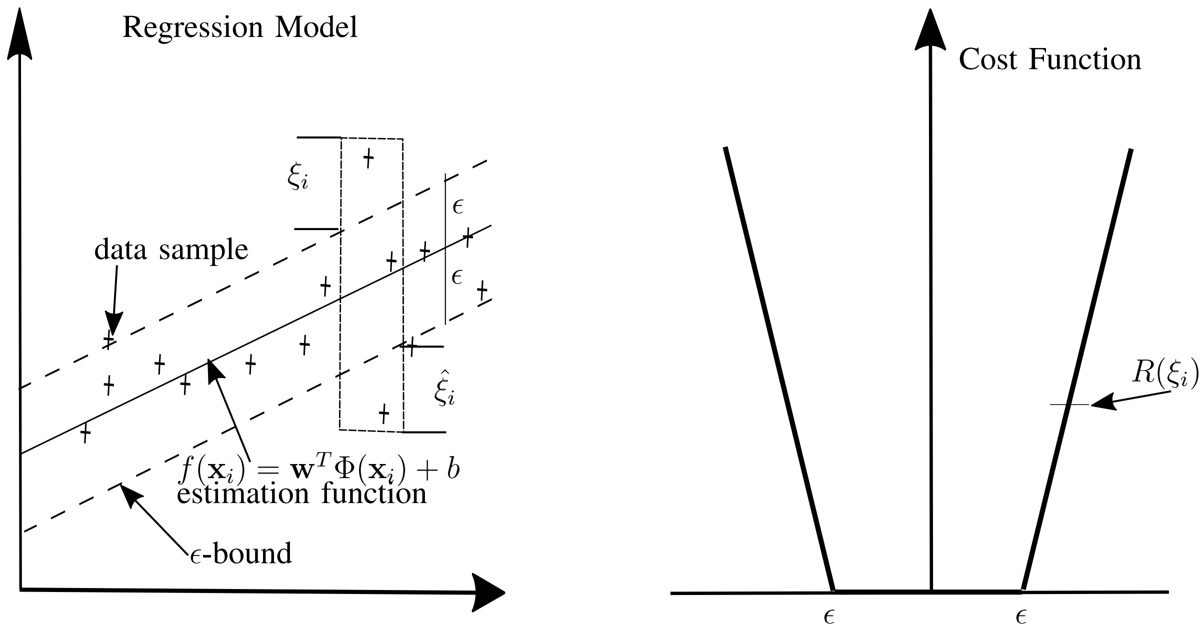

Fig. 1. ϵ-Support Vector Regression and Risk Functional

We present the primal optimization problem directly

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

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

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 solid line represents the estimation of the input-output relation, the area between two dash lines (ϵ-bound) is called ϵ-tube. Only those data samples located outside ϵ-tube contribute to the estimation error. Furthermore ξ_i and $\hat{\xi}_i$ denote slack variables that cope with additive noise of observations, $R(u)$ denotes the cost function. The simplest cost 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 distributed, 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.

In ϵ-SVR, the objective to exploit slack variables ξ_i and $\hat{\xi}_i$ is to compensate the influences from the outliers that exceed the ϵ-tube which are caused by noise. Therefore in ϵ-SVR, ξ_i and $\hat{\xi}_i$ are defined as

$$\xi_i = \max(0, y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b - \epsilon) \tag{3}$$

$$\hat{\xi}_i = \max(0, \mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i - \epsilon). \tag{4}$$

Because the distance between the estimations $\mathbf{w}^T\Phi(\mathbf{x}_i) + b$ and the observation $\mathbf{y}_i$ can only exceeds the ϵ-tube in one direction, therefore there is at most one of ξ_i and $\hat{\xi}_i$ can be non zero. That is $\xi_i\hat{\xi}_i = 0$.

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

$$\max_f \quad \prod_{i=1}^{L} P(y_i|f(\mathbf{x}_i)) = \prod_{i=1}^{L} P(f(\mathbf{x}_i) + \xi_i|f(\mathbf{x}_i)) = \prod_{i=1}^{L} Pr(\xi_i) = \prod_{i=1}^{L} Pr(y_i - f(\mathbf{x}_i)) \tag{5}$$

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

$$\sum_{i=1}^{L} \log(Pr(y_i - f(\mathbf{x}_i))), \tag{6}$$

maximizing (6) is equivalent to minimizing $-\sum_{i=1}^{L} \log(Pr(y_i - f(\mathbf{x}_i)))$, therefore (5) is equivalent to

$$\min_f - \sum_{i=1}^{L} \log(Pr(y_i - f(\mathbf{x}_i))), \tag{7}$$

Let $c(\mathbf{x}_i, y_i, f(\mathbf{x}_i))$ denotes the ith cost function, which is defined as

$$c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)) = -\log(P(y_i - f(\mathbf{x}_i))). \tag{8}$$

Thus (8) can be rewritten as

$$\min_f \quad \sum_{i=1}^{L} c(\mathbf{x}_i, y_i, f(\mathbf{x}_i)), \tag{9}$$

In ϵ-SVR, in order to provide more flexibility to precision control, Vapnik's ϵ-insensitive function, as shown in (10), is applied to (8).

$$|u|_\epsilon = \begin{cases} 0 & if \quad |u| < \epsilon \\ |u| - \epsilon & otherwise \end{cases} \tag{10}$$

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

$$\tilde{c}(y_i, \mathbf{x}_i, f(\mathbf{x}_i)) = -\log(Pr(|y_i - f(\mathbf{x}_i)|_\epsilon)), \tag{11}$$

Consider the cost function term in (2),

$$\sum_{i=1}^{L} R(\xi_i) + R(\hat{\xi}_i) = \sum_{i=1}^{L} \tilde{c}(y_i, \mathbf{x}_i, f(\mathbf{x}_i)) = \sum_{i=1}^{L} -\log(Pr(|y_i - f(\mathbf{x}_i)|_\epsilon)) \tag{12}$$

the cost function term is determined according to the distribution of noise.

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

$$\begin{aligned} L(\mathbf{w}, b, \xi_i, \hat{\xi}_i, \alpha_i, \hat{\alpha}_i, \eta_i, \hat{\eta}_i) &= \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{L}(R(\xi_i) + R(\hat{\xi}_i)) - \sum_{i=1}^{L}(\eta_i\xi_i + \hat{\eta}_i\hat{\xi}_i) \\ &+ \sum_{i=1}^{L}\alpha_i(y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \sum_{i=1}^{L}\hat{\alpha}_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i) \\ s.t. &\begin{cases} \eta_i, \hat{\eta}_i, \alpha_i, \hat{\alpha}_i \geq 0, i = 1, 2, \cdots L \\ \xi_i, \hat{\xi}_i \geq 0, i = 1, 2, \cdots L \end{cases} \end{aligned} \tag{13}$$

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 primal constrained optimization problem in (2) satisfies, are called Karush-Kuhn-Tucker (KKT) conditions. Here we elaborate a little further about how the dual objective problem is derived from KKT conditions.

Assume a constraint optimization problem is given by

$$\begin{aligned} &\min_{\mathbf{w}} \quad f(\mathbf{w}) \\ &s.t. \quad c_i(\mathbf{w}) \leq 0, i = 1, 2, \dots L, \end{aligned} \tag{14}$$

its Lagrange function is given by

$$L(\mathbf{w}, \mathbf{a}) = f(\mathbf{w}) + \sum_{i=1}^{L} a_i c_i(\mathbf{w}), \tag{15}$$

where $\mathbf{a} = [a_1, a_2, \dots, a_L]^T$ denotes the vector consist of Lagrange multipliers. Based on Theorem 6.21 in [26], a variable pair $[\bar{\mathbf{w}}, \bar{\mathbf{a}}]$ is the solution to (14) if and only if the following inequalities are satisfied

$$L(\mathbf{w}, \bar{\mathbf{a}}) \geq L(\bar{\mathbf{w}}, \bar{\mathbf{a}}) \geq L(\bar{\mathbf{w}}, \mathbf{a}) \tag{16}$$

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

$$\partial_{\mathbf{w}} L(\bar{\mathbf{w}}, \mathbf{a}) = \partial_{\mathbf{w}} f(\bar{\mathbf{w}}) + \sum_{i=1}^{L} a_i \partial_{\mathbf{w}} c_i(\bar{\mathbf{w}}) = 0, \tag{17}$$

$$\partial_{a_i} L(\bar{\mathbf{w}}, \bar{\mathbf{a}}) = c_i(\bar{\mathbf{w}}) \leq 0, i = 1, 2, \dots L \tag{18}$$

$$\bar{a}_i c_i(\bar{\mathbf{w}}) = 0, i = 1, 2, \dots L \tag{19}$$

In order to satisfy the first inequality in (16), (17) has to hold, applying (17) to (13), which are

$$\partial_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i)\Phi(\mathbf{x}_i) = 0 \tag{20}$$

$$\partial_{\xi_i} L = C_i R^{'}(\xi_i) - \eta_i - \alpha_i = 0, i = 1, 2, \cdots L \tag{21}$$

$$\partial_{\hat{\xi}_i} L = C_i R^{'}(\hat{\xi}_i) - \hat{\eta}_i - \hat{\alpha}_i = 0, i = 1, 2, \cdots L \tag{22}$$

$$\partial_b L = \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i) = 0 \tag{23}$$

Then by substituting (20)-(23) to (13), (13) can be rewritten as :

$$\theta(\alpha_i, \hat{\alpha}_i) = \frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\Phi^T(\mathbf{x}_j)\Phi(\mathbf{x}_i) + C\sum_{i=1}^{L}[R(\xi_i) + R(\hat{\xi}_i)] - \sum_{i=1}^{L}[(CR^{'}(\xi_i) - \alpha_i)\xi_i$$
$$+(CR^{'}(\hat{\xi}_i) - \hat{\alpha}_i)\hat{\xi}_i] + \sum_{i=1}^{L}\alpha_i[y_i - \sum_{j=1}^{L}(\alpha_j - \hat{\alpha}_j)\Phi^T(\mathbf{x}_j)\Phi(\mathbf{x}_i) - b - \epsilon - \xi_i] +$$
$$\sum_{i=1}^{L}\hat{\alpha}_i[\sum_{j=1}^{L}(\alpha_j - \hat{\alpha}_j)\Phi^T(\mathbf{x}_j)\Phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i] \tag{24}$$

notice in (23), $\sum_{i=1}^{L}(\alpha_i - \hat{\alpha}_i) = 0$, define $\tilde{R}(u) = R(u) - uR^{'}(u)$, (24) can be further simplified to

$$\theta(\alpha_i, \hat{\alpha}_i) = -\frac{1}{2}\sum_{i=1}^{L}\sum_{i=1}^{L}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i) + C\sum_{i=1}^{L}[(\tilde{R}(\xi_i) + \tilde{R}(\hat{\xi}_i)]$$
$$+\sum_{i=1}^{L}[(\alpha_i - \hat{\alpha}_i)y_i - (\alpha_i + \hat{\alpha}_i)\epsilon] \tag{25}$$

In order to satisfy the second inequality in (16), (18) and (19) have to hold. A $\mathbf{w}$ that satisfies the conditions in (18) is in the feasible region defined by inequality constraints as mentioned

(14). In ϵ-SVR, this condition is satisfied by making use of the slack variables ξ_i and $\hat{\xi}_i$ defined in (3) and (4). Define that $\tilde{\mathbf{w}}$ is in the feasible region and satisfies the conditions in (17), notice $\theta(\alpha_i, \hat{\alpha}_i)$ is equivalent to $L(\tilde{\mathbf{w}}, \mathbf{a})$, thus yielding the dual optimization problem, which is given by

$$\max_{\mathbf{a}} L(\tilde{\mathbf{w}}, \mathbf{a}) \equiv \max_{\alpha_i, \hat{\alpha}_i} \quad \theta(\alpha_i, \hat{\alpha}_i) = -\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\Phi^T(\mathbf{x}_j)\Phi(\mathbf{x}_i) + C\sum_{i=1}^{L}(\tilde{R}(\xi_i) + \tilde{R}(\hat{\xi}_i))$$
$$+\sum_{i=i}^{L}[(\alpha_i - \hat{\alpha}_i)y_i - (\alpha_i + \hat{\alpha}_i)\epsilon] \tag{26}$$

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 (26) can be written as

$$\max_{\mathbf{a}, \hat{\mathbf{a}}} \theta(\mathbf{a}, \hat{\mathbf{a}}) = -\frac{1}{2}(\mathbf{a} - \hat{\mathbf{a}})^T\mathbf{K}(\mathbf{a} - \hat{\mathbf{a}}) + (\mathbf{y} - \epsilon\mathbf{e})^T\mathbf{a} + (-\mathbf{y} - \epsilon\mathbf{e})^T\hat{\mathbf{a}} + C\mathbf{e}^T(\mathbf{R}_\xi + \mathbf{R}_{\hat{\xi}}), \tag{27}$$

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

$$[\mathbf{v}]_i = \begin{cases} 1 & i = 1, \ldots, l \\ -1 & i = l+1, \ldots, 2l \end{cases} \tag{28}$$

(27) can also be reformulate as

$$\max_{\mathbf{a}^*} \quad \theta(\mathbf{a}^*) = -\frac{1}{2}(\mathbf{a}^*)^T \left[\begin{smallmatrix}\mathbf{K} & -\mathbf{K}\\ -\mathbf{K} & \mathbf{K}\end{smallmatrix}\right] \mathbf{a}^{(*)} + [(\mathbf{y} - \epsilon)^T, (-\mathbf{y} - \epsilon)^T]\mathbf{a}^{(*)} + C\mathbf{e}^T(\mathbf{R}_\xi + \mathbf{R}_{\hat{\xi}}), \tag{29}$$

Condition in (19) is called KKT complementary condition, the value of $\sum_{i=1}^{L} \bar{a}_i c_i(\bar{\mathbf{w}})$ can be used to monitor the proximity of the current solution and the optimal solution. Thus it can be used as a stopping criterion. In the constraint optimization problem of (13), the KKT complementary conditions are given by

$$\begin{cases} \alpha_i(y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0, i = 1, 2 \cdots L \\ \hat{\alpha}_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i) = 0, i = 1, 2 \cdots L \\ (CR^{'}(\xi_i) - \alpha_i)\xi_i = 0, i = 1, 2, \ldots, L \\ (CR^{'}(\hat{\xi}_i) - \hat{\alpha}_i)\hat{\xi}_i = 0, i = 1, 2, \ldots, L \end{cases} \tag{30}$$

Based on the definitions of slack variables in (3) an (4), only when there is a outlier exists, ξ_i or $\hat{\xi}_i$ can be non zero, that is

$$\xi_i or \hat{\xi}_i = |\mathbf{y}_i - \mathbf{w}^T\Phi(\mathbf{x}_i)|_\epsilon, \tag{31}$$

because the distance of the estimation $\mathbf{w}^T\Phi(\mathbf{x}_i) + b$ and the observation y_i can only exceeds the ϵ-tube in one direction, as shown in Fig. 1. Therefore at most one of $(y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b - \epsilon - \xi_i)$ and $(\mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i)$ can be zero. In order to satisfy the KKT complementary conditions in (30), at least one of α_i and $\hat{\alpha}_i$ need to be zero, that is $\alpha_i\hat{\alpha}_i = 0$.

Therefore the complete KKT complementary conditions of ϵ-SVR is given by

$$\begin{cases} \alpha_i(y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0, i = 1, 2 \cdots L \\ \hat{\alpha}_i(\mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i) = 0, i = 1, 2 \cdots L \\ (CR^{'}(\xi_i) - \alpha_i)\xi_i = 0, i = 1, 2, \ldots, L \\ (CR^{'}(\hat{\xi}_i) - \hat{\alpha}_i)\hat{\xi}_i = 0, i = 1, 2, \ldots, L \\ \xi_i\hat{\xi}_i = 0, \alpha_i\hat{\alpha}_i = 0, i = 1, 2, \ldots L \end{cases} \tag{32}$$

III. DUAL CHANNEL COMPLEX SUPPORT VECTOR DETECTION FOR LARGE MIMO SYSTEM

A. 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.

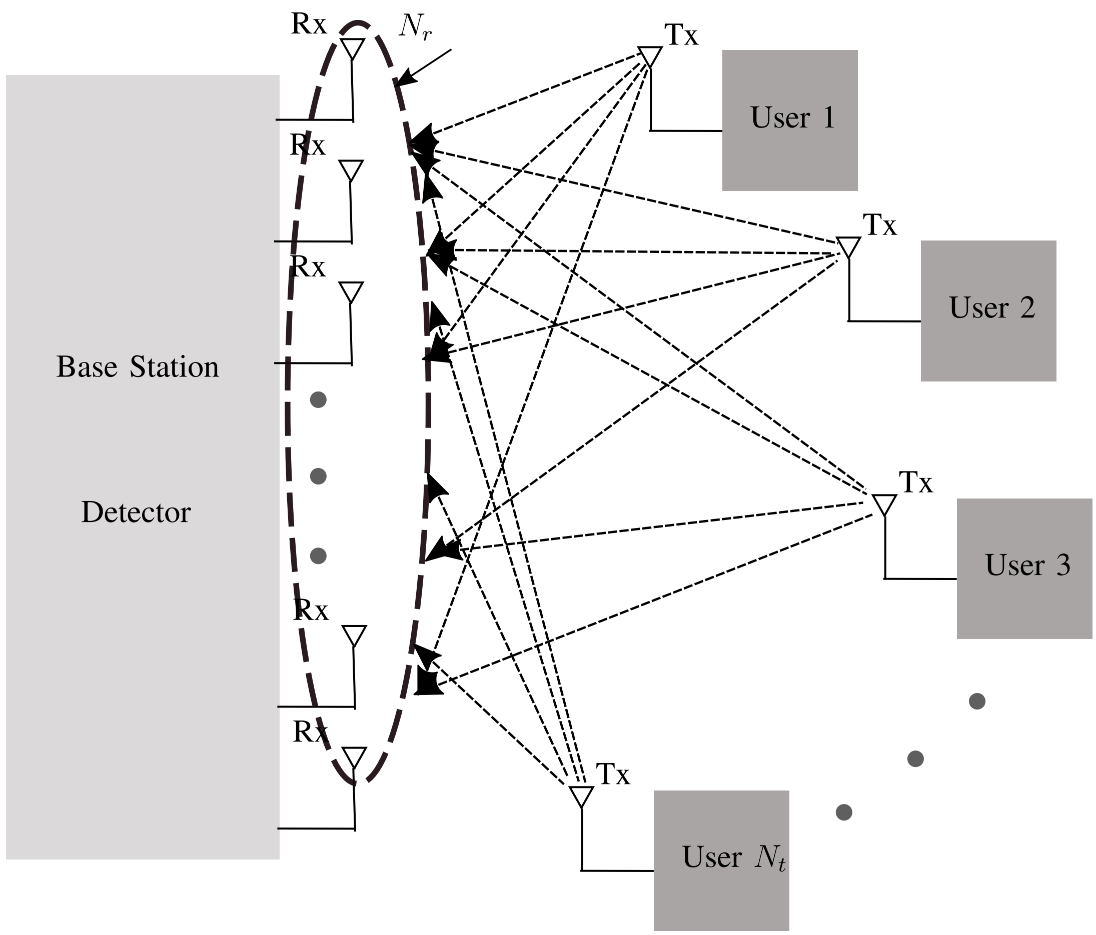

Fig. 2. Large MIMO uplink system

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:

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{33}$$

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 $\mathbf{H}$.

B. Complex Regression Model

Based on the discrete time model of a large MIMO uplink system of (33), our regression model is set such that the training data set at the detector are $(\mathbf{h}_1, y_1)(\mathbf{h}_2, y_2), \ldots, (\mathbf{h}_{N_r}, y_{N_r})$, where $\mathbf{h}_i$ denotes $\imath$th row of matrix $\mathbf{H}$. This yields a regression task without offset b in (1) :

$$y_i = f_{true}(\mathbf{h}_i) + n_i, i = 1, 2 \ldots L \tag{34}$$

$$f_{true}(\mathbf{h}_i) = \mathbf{h}_i \cdot \mathbf{s}, i = 1, 2, \ldots, L \tag{35}$$

$$\tag{36}$$

where $f_{true}()$ denotes the underlying true regression function, n_i denotes discrete sample of additive noise. In this regression problem, received symbols y_i are the observations, $\mathbf{h}_i$ are the input data samples, transmitted symbol vector $\mathbf{s}$ contains the regression coefficients. We employ complex support vector regression (CSVR) without offset term b. As shown in (17) the first KKT condition is satisfied by finding the saddle point $\bar{\mathbf{w}}$ by taking paritial derivatives of $L(\mathbf{w}, a_i)$ with respect to $\mathbf{w}$. Mathematical results of Wirtinger's calculus in Reproducing Kernel Hilbert Space (RKHS) are employed to calculate partial derivatives of Lagrangian function of CSVR in complex manner [39]. First we generalize our regression model by complex RKHS, Define a real RKHS $\mathcal{H}$, $<,>_{\mathcal{H}}$ denotes the corresponding inner product in $\mathcal{H}$. define complex RKHS $\mathbb{H} = \{f = f^r + \imath f^i, f^r, f^i \in \mathcal{H}\}$, $<,>_{\mathbb{H}}$ denotes the corresponding inner products in $\mathbb{H}$. Assume $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are complex vectors, μ, $\chi \in \mathbb{C}$, In general a complex Hilbert space $\mathbb{H}_g$ has the following properties

Property 1. $< \mathbf{x}, \mathbf{y} >_{\mathbb{H}_g} = (< \mathbf{y}, \mathbf{x} >_{\mathbb{H}_g})^*$

Property 2. $< \mu\mathbf{x} + \chi\mathbf{y}, \mathbf{z} >_{\mathbb{H}_g} = \mu < \mathbf{x}, \mathbf{z} >_{\mathbb{H}_g} + \chi < \mathbf{y}, \mathbf{z} >_{\mathbb{H}_g}$

Property 3. $< \mathbf{z}, \mu\mathbf{x} + \chi\mathbf{y} >_{\mathbb{H}_g} = \mu^* < \mathbf{z}, \mathbf{x} >_{\mathbb{H}_g} + \chi^* < \mathbf{z}, \mathbf{y} >_{\mathbb{H}_g}$

Lemma 1. *Assume* $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^V$, *define a real RKHS* $\mathcal{H}$ *that satisfy* $<\mathbf{a},\mathbf{b}>_{\mathcal{H}}=\mathbf{a}^T\mathbf{b}$. *For the corresponding complex RKHS* $\mathbb{H}=\{f=f^r+\imath f^i, f^r, f^i \in \mathcal{H}\}$, $\mathbf{m},\mathbf{l}\in\mathbb{C}^V$, $<\mathbf{m},\mathbf{l}^*>_{\mathbb{H}}=\mathbf{m}\cdot\mathbf{l}$

Proof. $\mathbf{g}$,$\mathbf{h} \in \mathbb{C}^V$, and $\mathbf{g}=\Re(\mathbf{g})+\imath\Im(\mathbf{g})$, $\mathbf{h}=\Re(\mathbf{h})+\imath\Im(\mathbf{h})$. From Property 1 and Property 3,

$$<\mathbf{g},\mathbf{h}>_{\mathbb{H}}=<\Re(g)+\imath\Im(g),\Re(h)+\imath\Im(h)>_{\mathbb{H}} \tag{37}$$

According to Property 2, (37) can be rewritten as

$$<\mathbf{g},\mathbf{h}>_{\mathbb{H}}=<\Re(\mathbf{g}),\Re(\mathbf{h})+\imath\Im(\mathbf{h})>_{\mathbb{H}}+\imath<\Im(\mathbf{g}),\Re(\mathbf{h})+\imath\Im(\mathbf{h})>_{\mathbb{H}}, \tag{38}$$

According to Property 3

$$\begin{aligned}&<\mathbf{g},\mathbf{h}>_{\mathbb{H}}=<\Re(\mathbf{g}),\Re(\mathbf{h})>_{\mathbb{H}}-\imath<\Re(\mathbf{g}),\Im(\mathbf{h})>_{\mathbb{H}}+\imath(<\Im(\mathbf{g}),\Re(\mathbf{h})>_{\mathbb{H}}-\imath<\Im(\mathbf{g}),\\&\Im(\mathbf{h})>_{\mathbb{H}})=<\Re(\mathbf{g}),\Re(\mathbf{h})>_{\mathbb{H}}+<\Im(\mathbf{g}),\Im(\mathbf{h})>_{\mathbb{H}}+\imath[<\Im(\mathbf{g}),\Re(\mathbf{h})>_{\mathbb{H}}-<\Re(\mathbf{g}),\\&\Im(\mathbf{h})>_{\mathbb{H}}]\end{aligned} \tag{39}$$

According to (39), if $\mathbf{f}_1,\mathbf{f}_2\in\mathcal{H}$, we have $<\mathbf{f}_1,\mathbf{f}_2>_{\mathbb{H}}=<\mathbf{f}_1,\mathbf{f}_2>_{\mathcal{H}}$, furthermore because $<\mathbf{f}_1,\mathbf{f}_2>_{\mathcal{H}}=\mathbf{f}_1^T\mathbf{f}_2$, (39) can be rewritten as

$$\begin{aligned}&<\mathbf{g},\mathbf{h}>_{\mathbb{H}}=<\Re(\mathbf{g}),\Re(\mathbf{h})>_{\mathcal{H}}+<\Im(\mathbf{g}),\Im(\mathbf{h})>_{\mathcal{H}}+\imath[<\Im(\mathbf{g}),\Re(\mathbf{h})>_{\mathcal{H}}-<\Re(\mathbf{g}),\Im(\mathbf{h})>_{\mathcal{H}}]=\\&\Re(\mathbf{g})^T\cdot\Re(\mathbf{h})+\Im(\mathbf{g})^T\cdot\Im(\mathbf{h})+\imath[\Im(\mathbf{g})^T\cdot\Re(\mathbf{h})-\Re(\mathbf{g})^T\cdot\Im(\mathbf{h})]=\mathbf{g}\cdot\mathbf{h}^*\end{aligned} \tag{40}$$

Therefore $<\mathbf{m},\mathbf{l}^*>_{\mathbb{H}}=\mathbf{m}^T\mathbf{l}$. ☐

Therefore we have $<\mathbf{h}_i,\mathbf{s}^*>_{\mathbb{H}}=\mathbf{h}_i\cdot\mathbf{s}$, For sake of simplicity we define $\mathbf{w}=\mathbf{s}^*$. Based on the same principle in the real case presented in section II, the primal optimization problem of large MIMO detection in complex RKHS $\mathbb{H}$ defined by $< \mathbf{m}, \mathbf{l} >_{\mathbb{H}} = \mathbf{m}^T \mathbf{l}^*$ can be formulated as

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

$$s.t. \begin{cases} \Re(y_i - < \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) \leq \epsilon + \xi_i^r, i = 1, 2, \ldots, N_r \\ \Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} - y_i) \leq \epsilon + \hat{\xi}_i^r, i = 1, 2, \ldots, N_r \\ \Im(y_i - < \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) \leq \epsilon + \xi_i^i, i = 1, 2, \ldots, N_r \\ \Im(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} - y_i) \leq \epsilon + \hat{\xi}_i^i, i = 1, 2, \ldots, N_r \\ \xi_i^r, \hat{\xi}_i^r, \xi_i^i, \hat{\xi}_i^i \geq 0, i = 1, 2, \ldots, N_r \end{cases} \tag{41}$$

Where $\Re(\cdot)$ and $\Im(\cdot)$ denote real and imaginary part of a complex variable and inequality constraints are set to real and imaginary part of the regression function separately. Let $\mathbf{K} = \mathbf{H}\mathbf{H}^H$ denote the kernel function, $\mathbf{K} = \Re(\mathbf{K}) + \imath \Im(\mathbf{K})$, where $\Re(\mathbf{K})$ and $\Im(\mathbf{K})$ denote matrices of corresponding real and imaginary parts. Similar to the Lagrange duality rational in section II-C, the Lagrange function associated with (41) is

$$\begin{aligned} L = & \frac{1}{2} ||\mathbf{w}||_{\mathbb{H}}^2 + C \sum_{i=1}^{N_r} [(R(\xi_i^r) + R(\hat{\xi}_i^r) + R(\xi_i^i) + R(\hat{\xi}_i^i))] - \sum_{i=1}^{N_r} (\eta_i \xi_i^r + \hat{\eta}_i \hat{\xi}_i^r + \tau_i \xi_i^i \\ & + \hat{\tau}_i \hat{\xi}_i^i) + \sum_{i=1}^{N_r} \alpha_i (\Re(y_i) - \Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - \epsilon - \xi_i^r) + \sum_{i=1}^{N_r} \hat{\alpha}_i (\Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - \Re(y_i) - \epsilon - \hat{\xi}_i^r) \\ & + \sum_{i=1}^{N_r} \beta_i (\Im(y_i) - \Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) - \epsilon - \xi_i^i) + \sum_{i=1}^{N_r} \hat{\beta}_i (\Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) - \Im(y_i) - \epsilon - \hat{\xi}_i^i) \\ & s.t. \begin{cases} \eta_i, \hat{\eta}_i, \tau_i, \hat{\tau}_i, \alpha_i, \hat{\alpha}_i, \beta_i, \hat{\beta}_i \geq 0, i = 1, 2, \ldots, N_r \\ \xi_i^r, \hat{\xi}_i^r, \xi_i^i, \hat{\xi}_i^i \geq 0, i = 1, 2, \ldots, N_r \end{cases} \end{aligned} \tag{42}$$

where $\eta_i, \hat{\eta}_i, \tau_i, \hat{\tau}_i, \alpha_i, \hat{\alpha}_i, \beta_i, \hat{\beta}_i$ are Lagrange multipliers. In order to calculate the partial derivations of L with respect to $\mathbf{w}$ which is defined in complex domain, the Wirtinger's calculus is exploited [39] , Therefore we have

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}^*} = \frac{1}{2}\mathbf{w} - \frac{1}{2}\sum_{i=1}^{N_r} \alpha_i \mathbf{h}_i + \frac{1}{2}\sum_{i=1}^{N_r} \hat{\alpha}_i \mathbf{h}_i + \frac{\imath}{2}(\sum_{i=1}^{N_r} \beta_i \mathbf{h}_i - \sum_{i=1}^{N_r} \hat{\beta}_i \mathbf{h}_i) = 0 \\ \Rightarrow \mathbf{w} = \sum_{i=1}^{N_r}(\alpha_i - \hat{\alpha}_i)\mathbf{h}_i - \imath\sum_{i=1}^{N_r}(\beta_i - \hat{\beta}_i)\mathbf{h}_i \\ \frac{\partial L}{\partial \xi_i^r} = CR^{'}(\xi_i^r) - \eta_i - \alpha_i = 0 \Rightarrow \eta_i = CR^{'}(\xi_i^r) - \alpha_i \\ \frac{\partial L}{\partial \hat{\xi}_i^r} = CR^{'}(\hat{\xi}_i^r) - \hat{\eta}_i - \hat{\alpha}_i = 0 \Rightarrow \hat{\eta}_i = CR^{'}(\hat{\xi}_i^r) - \hat{\alpha}_i \\ \frac{\partial L}{\partial \xi_i^i} = CR^{'}(\xi_i^i) - \tau_i - \beta_i = 0 \Rightarrow \tau_i = CR^{'}(\xi_i^i) - \beta_i \\ \frac{\partial L}{\partial \hat{\xi}_i^i} = CR^{'}(\hat{\xi}_i^i) - \hat{\tau}_i - \hat{\beta}_i = 0 \Rightarrow \hat{\tau}_i = CR^{'}(\hat{\xi}_i^i) - \hat{\beta}_i \end{cases} \tag{43}$$

Based on (43), we have

$$\begin{aligned} &< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} = \sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} + i\sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} \\ &= \sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\mathbf{h}_i\mathbf{h}_j^H + i\sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\mathbf{h}_i\mathbf{h}_j^H \\ &= \sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\Re(\mathbf{K})_{ij} - \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\Im(\mathbf{K})_{ij} + i(\sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\Im(\mathbf{K})_{ij} + \\ &\sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\Re(\mathbf{K})_{ij}), \end{aligned} \tag{44}$$

$$\begin{aligned} ||\mathbf{w}||_{\mathbb{H}}^2 &= \sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_i - \hat{\alpha}_i)\mathbf{h}_i\mathbf{h}_j^H + \sum_{i,j=1}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_i)\mathbf{h}_i\mathbf{h}_j^H \\ &+ i(\sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j)\mathbf{h}_i\mathbf{h}_j^H - \sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j)\mathbf{h}_j\mathbf{h}_i^H) \end{aligned} \tag{45}$$

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

$$\sum_{i,j}^{L} r_i r_j \Im(\mathbf{K})_{ij} = -\sum_{i,j}^{L} r_i r_j \Im(\mathbf{K})_{ji} = -\sum_{i,j}^{L} r_i r_j \Im(\mathbf{K})_{ij}, \tag{46}$$

Therefore

$$\sum_{i,j}^{l} r_i r_j \Im(\mathbf{K})_{ij} = 0, \tag{47}$$

Based on (47), (45) can be changed to

$$||\mathbf{w}||^2_{\mathbb{H}} = \sum_{i,j=1}^{N_r}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\Re(\mathbf{K})_{ij}+\sum_{i,j=1}^{N_r}(\beta_i-\hat{\beta}_i)(\beta_j-\hat{\beta}_j)\Re(\mathbf{K})_{ij}-2\sum_{i,j=1}^{N_r}(\alpha_i-\hat{\alpha}_i)(\beta_j-\hat{\beta}_j)\Im(\mathbf{K})_{ij}. \tag{48}$$

Apply (43), (44), (47) and (48) to (42), we have

$$\begin{aligned}
L = {} & \frac{1}{2}[\sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\Re(\mathbf{K})_{ij} + \sum_{i,j=1}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)\Re(\mathbf{K})_{ij} - 2\sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j)\Im(\mathbf{K})_{ij}]+ \\
& C\sum_{i=1}^{N_r}[R(\xi_i^r) + R(\hat{\xi}_i^r) + R(\xi_i^i) + R(\hat{\xi}_i^i)] - \sum_{i=1}^{N_r}((CR^{'}(\xi_i^r) - \alpha_i)\xi_i^r + (CR^{'}(\hat{\xi}_i^r) - \hat{\alpha}_i)\hat{\xi}_i^r + (CR^{'} - \beta_i)\xi_i^i \\
& +(CR^{'}(\hat{\xi}_i) - \hat{\beta}_i)\hat{\xi}_i^i) + \sum_{i=1}^{N_r}\alpha_i(\Re(y_i) - (\sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\Re(\mathbf{K})_{ij} - \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\Im(\mathbf{K})_{ij}) - \epsilon - \xi_i^r) + \sum_{i=1}^{N_r}\hat{\alpha}_i \\
& ((\sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\Re(\mathbf{K})_{ij} - \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\Im(\mathbf{K})_{ij}) - \Re(y_i) - \epsilon - \hat{\xi}_i^r) + \sum_{i=1}^{N_r}\beta_i(\Im(y_i) - (\sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j) \\
& \Im(\mathbf{K})_{ij} + \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\Re(\mathbf{K})_{ij}) - \epsilon - \xi_i^i) + \sum_{i=1}^{N_r}\hat{\beta}_i((\sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\Im(\mathbf{K})_{ij} + \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\Re(\mathbf{K})_{ij}) - \Im(y_i)- \\
& \epsilon - \hat{\xi}_i^i)
\end{aligned} \tag{49}$$

Then simplify (49), we obtain

$$L \equiv \theta = -\frac{1}{2}[\sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\Re(\mathbf{K})_{ij} + \sum_{i,j=1}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)\Re(\mathbf{K})_{ij}] + C\sum_{i=1}^{N_r}[R(\xi_i^r) - \xi_i^r R^{'}(\xi_i^r)$$
$$+R(\hat{\xi}_i^r) - \hat{\xi}_i^r R^{'}(\hat{\xi}_i^r) + R(\xi_i^i) - \xi_i R^{'}(\xi_i^i) + R(\hat{\xi}_i^i) - \hat{\xi}_i^i R^{'}(\hat{\xi}_i^i)] + \sum_{i=1}^{N_r}[\Re(y_i)(\alpha_i - \hat{\alpha}_i) + \Im(y_i)(\beta_i - \hat{\beta}_i)] -$$
$$\epsilon\sum_{i=1}^{N_r}(\alpha_i + \hat{\alpha}_i + \beta_i + \hat{\beta}_i) \tag{50}$$

define $\tilde{R}(u) = R(u) - uR^{'}(u)$. Similar to (26), The dual optimization problem of a complex MIMO system is given by

$$\max_{\alpha_i,\hat{\alpha}_i,\beta_i,\hat{\beta}_i} \quad \theta = -\frac{1}{2}[\sum_{i,j}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{ij}^r + \sum_{i,j}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)\mathbf{K}_{ij}^r]$$
$$-\sum_{i}^{N_r}(\alpha_i + \hat{\alpha}_i + \beta + \hat{\beta}_i)\epsilon + [\sum_{i=1}^{N_r}(\alpha_i - \hat{\alpha}_i)Re(y_i) + \sum_{i=1}^{N_r}(\beta_i - \hat{\beta}_i)Im(y_i)]$$
$$+C\sum_{i}^{N_r}(\tilde{R}(\xi_i^r) + \tilde{R}(\hat{\xi}_i^r) + \tilde{R}(\xi_i^i) + \tilde{R}(\hat{\xi}_i^i))$$
$$\begin{cases} 0 \le \alpha_i(\hat{\alpha}_i) \le C\tilde{R}(\xi_i^r)(\tilde{R}(\hat{\xi}_i^r)), i = 1, 2, \ldots, L \\ 0 \le \beta_i(\hat{\beta}_i) \le C\tilde{R}(\xi_i^i)(\tilde{R}(\hat{\xi}_i^i)), i = 1, 2, \ldots, L \\ \xi_i^r(\hat{\xi}_i^r) \ge 0, i = 1, 2, \ldots, L \\ \xi_i^i(\hat{\xi}_i^i) \ge 0, i = 1, 2, \ldots, L \end{cases} \tag{51}$$

Notice in (51), there is no correlation term between $\alpha_i, \hat{\alpha}_i$ and $\beta_i, \hat{\beta}_i$, therefore, (51) can be divided into two independent regression tasks, which are given by

$$\max_{\alpha_i,\hat{\alpha}_i} \quad \theta^r = -\frac{1}{2}\sum_{i,j}^{N_r}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\Re(\mathbf{K})_{ij} - \sum_{i=1}^{N_r}(\alpha_i+\hat{\alpha}_i)\epsilon + \sum_{i=1}^{N_r}(\alpha_i-\hat{\alpha}_i)Re(y_i) + C\sum_{i=1}^{N_r}(\tilde{R}(\xi_i^r)$$
$$+\tilde{R}(\hat{\xi}_i^r))$$
$$\begin{cases} 0 \leq \alpha_i(\hat{\alpha}_i) \leq C\tilde{R}(\xi_i^r)(\tilde{R}(\hat{\xi}_i^r)), i=1,2,\ldots,L \\ \xi_i^r(\hat{\xi}_i^r) \geq 0, i=1,2,\ldots,L \end{cases} \tag{52}$$

and

$$\max_{\beta_i,\hat{\beta}_i} \quad \theta^i = -\frac{1}{2}\sum_{i,j}^{N_r}(\beta_i-\hat{\beta}_i)(\beta_j-\hat{\beta}_j)\Re(\mathbf{K})_{ij} - \sum_{i=1}^{N_r}(\beta_i+\hat{\beta}_i)\epsilon + \sum_{i=1}^{N_r}(\beta_i-\hat{\beta}_i)Im(y_i) + C\sum_{i=1}^{N_r}(\tilde{R}(\xi_i^i)$$
$$+\tilde{R}(\hat{\xi}_i^i))$$
$$\begin{cases} 0 \leq \beta_i(\hat{\beta}_i) \leq C\tilde{R}(\xi_i^i)(\tilde{R}(\hat{\xi}_i^i)), i=1,2,\ldots,L \\ \xi_i^i(\hat{\xi}_i^i) \geq 0, i=1,2,\ldots,L \end{cases} \tag{53}$$

Let $\mathbf{a} = [\alpha_1, \alpha_2, \ldots, \alpha_{N_r}]^T$, $\hat{\mathbf{a}} = [\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{N_r}]^T$, $\mathbf{b} = [\beta_1, \beta_2, \ldots, \beta_{N_r}]^T$,$\hat{\mathbf{b}} = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{N_r}]^T$, $\mathbf{R}^r = [\tilde{R}(\xi_1^r) + \tilde{R}(\hat{\xi}_1^r), \tilde{R}(\xi_2^r) + \tilde{R}(\hat{\xi}_2^r), \ldots, \tilde{R}(\xi_{N_r}^r) + \tilde{R}(\hat{\xi}_{N_r}^r)]^T$, $\mathbf{R}^i = [\tilde{R}(\xi_1^i) + \tilde{R}(\hat{\xi}_1^i), \tilde{R}(\xi_2^i) + \tilde{R}(\hat{\xi}_2^i), \ldots, \tilde{R}(\xi_{N_r}^i) + \tilde{R}(\hat{\xi}_{N_r}^i)]^T$, $\mathbf{e} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{N_r}$. The alternate form can be written as

$$\max_{\mathbf{a},\hat{\mathbf{a}}} \quad \theta^r = -\frac{1}{2}(\mathbf{a}-\hat{\mathbf{a}})^T\Re(\mathbf{K})(\mathbf{a}-\hat{\mathbf{a}}) + \Re(\mathbf{y})^T(\mathbf{a}-\hat{\mathbf{a}}) - \epsilon(\mathbf{e}^T(\mathbf{a}+\hat{\mathbf{a}})) + C(\mathbf{e}^T\mathbf{R}^r)$$
$$\begin{cases} 0 \leq \alpha_i(\hat{\alpha}_i) \leq C\tilde{R}(\xi_i^r)(\tilde{R}(\hat{\xi}_i^r)), i=1,2,\ldots,N_r \\ \xi_i^r(\hat{\xi}_i^r) \geq 0, i=1,2,\ldots,N_r \end{cases} \tag{54}$$

$$
\max_{\mathbf{b},\hat{\mathbf{b}}} \quad \theta^i = -\frac{1}{2}(\mathbf{b}-\hat{\mathbf{b}})^T \Re(\mathbf{K})(\mathbf{b}-\hat{\mathbf{b}}) + \Im(\mathbf{y})^T(\mathbf{b}-\hat{\mathbf{b}}) - \epsilon(\mathbf{e}^T(\mathbf{b}+\hat{\mathbf{b}})) + C(\mathbf{e}^T\mathbf{R}^i)
$$

$$
\begin{cases} 0 \leq \beta_i(\hat{\beta}_i) \leq C\tilde{R}(\xi_i^i)(\tilde{R}(\hat{\xi}_i^i)), i = 1, 2, \ldots, N_r \\ \xi_i^i(\hat{\xi}_i^i) \geq 0, i = 1, 2, \ldots, N_r \end{cases} \tag{55}
$$

Observe that solving (54) and (55) are equivalent to solving two independent real Support vector regression task (dual channel), where only the real part of the kernel matrix is required for each channel. This can provide computational advantages for solving the dual optimization problems.

IV. Work Set Selection and Solver

As we can see, a quadratic optimization problem is formulated in (51). 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 a large data set.

Decomposition methods are a set of efficient algorithms that can help to ease this difficulty. Decomposition methods work iteratively, and are based on choosing a subset of optimization variables S (named work set) to be updated in each iteration step, while keeping the rest variables N fixed. Sequential Minimal Optimization (SMO) is an extreme case of decomposition method, where the work set size is 2, and an analytic quadratic programming (QP) step instead of numerical QP step can be performed in each iteration.

Because (52) and (53) have the same structure, in this section we discuss the real part only.

By dividing the variables into work set S and fixed set N. Therefore, the vector $\mathbf{a}$ in (54) can be divided into two sub vectors $[\mathbf{a}_S, \mathbf{a}_N]$. thereafter $\mathbf{a}_S$ denotes a vector that consists of the components of $\mathbf{a}$ that belongs to set S, the same modifications are applied to $\hat{\mathbf{a}}$ and $\mathbf{y}$. Thus (54) can be rewritten as:

$$\max_{\mathbf{a}_S,\hat{\mathbf{a}}_S,\mathbf{a}_N,\hat{\mathbf{a}}_N} \theta = -\frac{1}{2}[(\mathbf{a}_S - \hat{\mathbf{a}}_S)^T, (\mathbf{a}_N - \hat{\mathbf{a}}_N)^T] \begin{bmatrix} \Re(\mathbf{K})_{SS} & \Re(\mathbf{K})_{SN} \\ \Re(\mathbf{K})_{NS} & \Re(\mathbf{K})_{NN} \end{bmatrix} \begin{bmatrix} (\mathbf{a}_S - \hat{\mathbf{a}}_S) \\ (\mathbf{a}_N - \hat{\mathbf{a}}_N) \end{bmatrix} + [\Re(\mathbf{y}_S)^T, \Re(\mathbf{y}_N)^T] \begin{bmatrix} (\mathbf{a}_S - \hat{\mathbf{a}}_S) \\ (\mathbf{a}_N - \hat{\mathbf{a}}_N) \end{bmatrix} -$$

$$\epsilon[\mathbf{e}_S^T, \mathbf{e}_N^T] \begin{bmatrix} (\mathbf{a}_S + \hat{\mathbf{a}}_S) \\ (\mathbf{a}_N + \hat{\mathbf{a}}_N) \end{bmatrix} + C(\mathbf{e}^T, \mathbf{R}^r)$$

$$\begin{cases} 0 \le [\mathbf{a}]_i([\hat{\mathbf{a}}]_i) \le C\tilde{R}(\xi_i^r)(\tilde{R}(\hat{\xi}_i^r)), i = 1, 2, \ldots, N_r \\ \xi_i^r(\hat{\xi}_i^r) \ge 0, i = 1, 2, \ldots, N_r \end{cases} \tag{56}$$

Where $\begin{bmatrix} \Re(\mathbf{K})_{SS} & \Re(\mathbf{K})_{SN} \\ \Re(\mathbf{K})_{NS} & \Re(\mathbf{K})_{NN} \end{bmatrix}$ is a permutation of $\Re(\mathbf{K})$, where $\Re(\mathbf{K})_{SN}$ denotes the sub matrix that consist of the components in $\Re(\mathbf{K})$ that belong to both of the rows corresponding to set S and columns corresponding to N, Notice that $\mathbf{K}$ is a Hermitian matrix, thus $\Re(\mathbf{K})_{SN} = \Re(\mathbf{K})_{NS}$, therefore, (56) can be rewritten as

$$\max_{\mathbf{a},\hat{\mathbf{a}}} \quad \theta^r = -\frac{1}{2}[(\mathbf{a}_S - \hat{\mathbf{a}}_S)^T \Re(\mathbf{K})_{SS}(\mathbf{a}_S - \hat{\mathbf{a}}_S) + 2(\mathbf{a}_N - \hat{\mathbf{a}}_N)^T \Re(\mathbf{K})_{NS}(\mathbf{a}_S - \hat{\mathbf{a}}_S)] + \Re(\mathbf{y}_S^T)(\mathbf{a}_S - \hat{\mathbf{a}}_S) -$$

$$\epsilon(\mathbf{e}^T(\mathbf{a}_S + \hat{\mathbf{a}}_S)) - \frac{1}{2}(\mathbf{a}_N - \hat{\mathbf{a}}_N)^T \Re(\mathbf{K})_{NN}(\mathbf{a}_N - \hat{\mathbf{a}}_N) + \Re(\mathbf{y}_N^T)(\mathbf{a}_N - \hat{\mathbf{a}}_N) - \epsilon(\mathbf{e}^T(\mathbf{a}_N + \hat{\mathbf{a}}_N))$$

$$+C(\mathbf{e}^T(\mathbf{R}^r),$$

$$\begin{cases} 0 \le [\mathbf{a}]_i([\hat{\mathbf{a}}]_i) \le C\tilde{R}(\xi_i^r)(\tilde{R}(\hat{\xi}_i^r)), i = 1, 2, \ldots, N_r \\ \xi_i^r(\hat{\xi}_i^r) \ge 0, i = 1, 2, \ldots, N_r \end{cases} \tag{57}$$

In each iteration, in (57), $\mathbf{a}_N$ is fixed and only the sub QP problem that related to $\mathbf{a}_S$ is considered i.e

$$\max_{\mathbf{a}_S,\hat{\mathbf{a}}_S} \quad \theta_S^r = -\frac{1}{2}[(\mathbf{a}_S-\hat{\mathbf{a}}_S)^T\Re(\mathbf{K})_{SS}(\mathbf{a}_S-\hat{\mathbf{a}}_S)] + [Re(\mathbf{y}_S^T) - (\mathbf{a}_N-\hat{\mathbf{a}}_N)^T\mathbf{K}_{NS}^r](\mathbf{a}_S-\hat{\mathbf{a}}_S)-$$

$$\epsilon < \mathbf{e}^T, (\mathbf{a}_S+\hat{\mathbf{a}}_S)>,$$

$$\begin{cases} 0 \leq [\mathbf{a}_S]_i([\hat{\mathbf{a}}_S]_i) \leq C\tilde{R}(\xi_i^r)(\tilde{R}(\hat{\xi}_i^r)), i \in S \\ \xi_i^r(\hat{\xi}_i^r) \geq 0, i \in S \end{cases} \tag{58}$$

A proper work set selection strategy is required so that speed and performance requirement can be guaranteed. One idea is to perform update to optimization variables that violate the KKT complementary conditions, which, based on (32)

$$\begin{cases} (C\tilde{R}(\xi_i^r)-\alpha_i)\xi_i^r = 0, i = 1,2,\ldots,L \\ (C\tilde{R}(\hat{\xi}_i^r)-\hat{\alpha}_i)\hat{\xi}_i^r = 0, i = 1,2,\ldots,L \\ \alpha_i(\Re(\mathbf{y}_i)- <\mathbf{h}_i,\mathbf{w}>_{\mathbb{H}} -\epsilon-\xi_i^r) = 0, i = 1,2,\ldots,L \\ \hat{\alpha}_i(<\mathbf{h}_i,\mathbf{w}>_{\mathbb{H}} -\Re(\mathbf{y}_i)-\epsilon-\hat{\xi}_i^r) = 0, i = 1,2,\ldots,L \\ \xi_i\hat{\xi}_i = 0, \alpha_i\hat{\alpha}_i = 0, i = 1,2,\ldots,L \end{cases} \tag{59}$$

According to Osuna's theorem [35], the final convergence can be guaranteed. In SMO algorithm, heuristic methods are used to find a work set of size two in order to accelerate the decomposition process [34]. The heuristic method first searches among the non-bound variables (that is $0 < \alpha_i < C\tilde{R}(\xi_i^r)$ and $0 < \hat{\alpha}_i < C\tilde{R}(\hat{\xi}_i)$), which are more likely to violate the complementary KKT conditions, Then searching the whole dual variable set. The second dual variable that can maximize optimization step size is chosen. An approximate step size is used as evaluator for sake of reducing computational cost.

Another idea for work set selection is to choose the optimization variables whose update can provide the maximum improvement to the sub dual objective functions. The improvement of the sub dual objective function is

$$\bigtriangledown\theta_S^r = \theta_S^r((\mathbf{a}_S + \Delta_S), (\hat{\mathbf{a}}_S + \hat{\Delta}_S)) - \theta_S^r(\mathbf{a}_S, \hat{\mathbf{a}}_S), \tag{60}$$

where $\Delta_S = \mathbf{a}_S^{new} - \mathbf{a}_S$, $\hat{\Delta}_S = \hat{\mathbf{a}}_S^{new} - \hat{\mathbf{a}}_S$, $\mathbf{a}_S^{new}$ and $\hat{\mathbf{a}}_S^{new}$ are the updates of $\mathbf{a}_S$ and $\hat{\mathbf{a}}_S$. $\bigtriangledown\theta_S^r$ in (60), based on (58), can be written as

$$\begin{aligned}
&\bigtriangledown\theta_S^r = -\frac{1}{2}[(\Delta_S - \hat{\Delta}_S)^T \Re(\mathbf{K})_{SS}(\Delta_S - \hat{\Delta}_S) + 2(\mathbf{a}_S - \hat{\mathbf{a}}_S)^T \Re(\mathbf{K})_{SS}(\Delta_S - \hat{\Delta}_S)] + [\Re(\mathbf{y}_S^T) - (\mathbf{a}_N - \hat{\mathbf{a}}_N)^T \\
&\Re(\mathbf{K})_{NS}](\Delta_S - \hat{\Delta}_S) - \epsilon\mathbf{e}_S^T(\Delta_S + \hat{\Delta}_S) = -\frac{1}{2}(\Delta_S - \hat{\Delta}_S)^T \Re(\mathbf{K})_{SS}(\Delta_S - \hat{\Delta}_S) + \{\Re(\mathbf{y}_S^T) - [(\mathbf{a}_S - \hat{\mathbf{a}}_S)^T \\
&\Re(\mathbf{K})_{SS} + (\mathbf{a}_N - \hat{\mathbf{a}}_N)^T \Re(\mathbf{K})_{NS}]\}(\Delta_S - \hat{\Delta}_S) - \epsilon\mathbf{e}_S^T(\Delta_S + \hat{\Delta}_S),
\end{aligned} \tag{61}$$

where we have

$$\begin{aligned}
&(\mathbf{a}_S - \hat{\mathbf{a}}_S)^T \Re(\mathbf{K})_{SS} + (\mathbf{a}_N - \hat{\mathbf{a}}_N)^T \Re(\mathbf{K})_{NS} = [(\mathbf{a}_S - \hat{\mathbf{a}}_S)^T, (\mathbf{a}_N - \hat{\mathbf{a}}_N)^T] \begin{bmatrix} \Re(\mathbf{K})_{SS} \\ \Re(\mathbf{K})_{NS} \end{bmatrix} = \\
&(\mathbf{a} - \hat{\mathbf{a}})^T \Re(\mathbf{K})_S,
\end{aligned} \tag{62}$$

where $\Re(\mathbf{K})_S \in \mathbb{R}^{N_r \times S}$ denotes the matrix constructed by all the columns corresponding to the work set S. Therefore, (61) can be rewritten as

$$\begin{aligned}
&\bigtriangledown\theta_S^r = -\frac{1}{2}(\Delta_S - \hat{\Delta}_S)^T \Re(\mathbf{K})_{SS}(\Delta_S - \hat{\Delta}_S) + [\Re(\mathbf{y}_S^T) - (\mathbf{a} - \hat{\mathbf{a}})^T \Re(\mathbf{K})_S](\Delta_S - \hat{\Delta}_S) - \epsilon\mathbf{e}_S^T \\
&(\Delta_S + \hat{\Delta}_S),
\end{aligned} \tag{63}$$

define intermediate variable vector $\Phi^r, \Phi^i \in \mathbb{R}^{N_r}$, $\Phi^r = \Re(\mathbf{y}) - \Re(\mathbf{K})(\mathbf{a} - \hat{\mathbf{a}})$ and $\Phi^i = \Im(\mathbf{y}) -$

$\Re(\mathbf{K})(\mathbf{b}-\hat{\mathbf{b}})$. Let Φ_S denote the vector that consists of the components in Φ that corresponding to S. Thus (63) can be rewritten as

$$\bigtriangledown\theta_S^r = -\frac{1}{2}(\Delta_S - \hat{\Delta}_S)^T \Re(\mathbf{K})_{SS}(\Delta_S - \hat{\Delta}_S) + (\Phi_S^r)^T(\Delta_S - \hat{\Delta}_S) - \epsilon \mathbf{e}_S^T (\Delta_S + \hat{\Delta}_S) \tag{64}$$

Because the offset term is omitted in the large MIMO regression model, therefore different from SMO type algorithms, there is no linear equality constraint, which is inducted by offset b, as shown in (23). Therefore the size of work set can be less than two. However, more efficient work set selection strategies based on maximal sub dual objective gain selection, is proposed in [36]. A work set of size two is updated in each iteration. The computational cost is reduced while maintaining the comparable performance with the model with offset term. In the CSVR large MIMO detector, the work set selection strategy in [36] is exploited. Basically, in each iteration this strategy uses the combined single variable searchings, whose searching time is $O(n)$, where n is the number of data samples, to determine a work set of size two. On the one hand, this strategy can find a work set of size two, whose update approximately maximize the gain of the dual sub objective function. On the other hand, the brute-force searching for the work size of two, whose update can maximize the gain of the sub dual objective function, requires $O(n^2)$ searching time. The former strategy requires as few iterations as the latter one, since in each iteration, the latter strategy is more computationally expensive, the former one can enjoy significantly faster running speed.

A. Single Direction (1-D) Solver

Recall the KKT complementary conditions

$$(C\tilde{R}(\xi_i^r) - \alpha_i)\xi_i^r = 0, i = 1, 2, \ldots, N_r \tag{65}$$

$$(C\tilde{R}(\hat{\xi}_i^r) - \hat{\alpha}_i)\hat{\xi}_i^r = 0, i = 1, 2, \ldots, N_r \tag{66}$$

$$\alpha_i(\Re(\mathbf{y}_i) - < \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} - \epsilon - \xi_i^r) = 0, i = 1, 2, \ldots, N_r \tag{67}$$

$$\hat{\alpha}_i(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} - \Re(\mathbf{y}_i) - \epsilon - \hat{\xi}_i^r) = 0, i = 1, 2, \ldots, N_r \tag{68}$$

$$0 \leq \alpha_i(\hat{\alpha}_i) \leq R^{'}(\xi_i^r)(R^{'}(\hat{\xi}_i)), i = 1, 2, \ldots, N_r \tag{69}$$

Recall the discussions of slack variables by (3) and (4) in section II-C. we have $\xi_i^r \hat{\xi}_i^r = 0$. similar to the definitions in (31), we have

$$\xi_i^r or \hat{\xi}_i^r = |\Re(\mathbf{y}_i) - \Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})|_\epsilon, \tag{70}$$

therefore in (67) and (68), there is at least one of $(\Re(\mathbf{y}_i) - < \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} - \epsilon - \xi_i^r)$ and $(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} - \Re(\mathbf{y}_i) - \epsilon - \hat{\xi}_i^r)$ can be non zero, therefore in order to satisfy equalities (67) and (68), at least one of α_i and $\hat{\alpha}_i$ need to be zero, that is $\alpha_i \hat{\alpha}_i = 0$.

Hence we can substitute $\lambda_i = \alpha_i - \hat{\alpha}_i$ and $|\lambda_i| = \alpha_i + \hat{\alpha}_i$, therefore the update unit is changed to the single optimization variable λ_i rather than the pair α_i and $\hat{\alpha}_i$.

We first introduce 1-D work set selection strategy in which single optimization variable whose update can maximize the gain of the sub dual objective function is chosen as the work set in each iteration. Define $\Lambda = [\lambda_1, \lambda_2, \ldots, \lambda_{N_r}]^T$, $\sigma_i = \lambda_i^{new} - \lambda_i$, $\Sigma = [\sigma_1, \sigma_2, \ldots, \sigma_{N_r}]^T$. In (58), substitute $\mathbf{a}, \hat{\mathbf{a}}$ by Λ and in (64), substitute Δ by Σ. Let Λ_S and Σ_S denote the vector that consist of the components in Λ and Σ that belong to work set S, the sub dual objective function in (58) and its gain (64) can be rewritten as

$$\max_{\Lambda_S} \quad \theta_S^r = -\frac{1}{2}[\Lambda_S^T \Re(\mathbf{K})_{SS} \Lambda_S] + [\Re(\mathbf{y}_S^T) - \Lambda_N^T \Re(\mathbf{K})_{NS}]\Lambda_S - \epsilon < \mathbf{e}_S^T, |\Lambda_S| >, \tag{71}$$

$$\bigtriangledown \theta_S^r = -\frac{1}{2}\Sigma_S^T \Re(\mathbf{K})_{SS} \Sigma_S + (\Phi_S^r)^T \Sigma_S - \epsilon < \mathbf{e}_S^T, |\Lambda_S^{new}| - |\Lambda_S| >, \tag{72}$$

In 1-D solver, $|S| = 1$, based on (71), the sub dual objective function corresponding to the mth optimization variable $\tilde{\lambda}_m^{new}$, $m = 1, 2, \ldots, N_r$ can be written as

$$\max_{\tilde{\lambda}_m^{new}} \quad \theta_m^r = -\frac{1}{2}(\tilde{\lambda}_m^{new})^2 \Re(\mathbf{K})_{mm} + [\Re(\mathbf{y}_m) - \sum_{j \neq m}^{N_r} \Re(\mathbf{K})_{mj} \lambda_j] \tilde{\lambda}_m^{new} - \epsilon(|\tilde{\lambda}_m^{new}|), \tag{73}$$

where $\tilde{\lambda}_m^{new}$ is the update of λ_m. Based on the definition of Φ in (64), $\Phi_i^r = \Re(\mathbf{y}_i) - \sum_{j=1}^{N_r} \lambda_j^r \Re(\mathbf{K})_{ij}$, similarly, as to dual variable λ_i^i in the imaginary channel, $\Phi_i^i = \Im(y_i) - \sum_{j=1}^{N_r} \lambda_j^i \Re(\mathbf{K})_{ij}$. Thereafter, for sake of brevity, we use λ_i instead of λ_i^r. Take the partial derivative of θ_m^r respect to $\tilde{\lambda}_m^{new}$, we have

$$\frac{\partial \theta_m^r}{\partial \tilde{\lambda}_m^{new}} = -\tilde{\lambda}_m^{new} \Re(\mathbf{K})_{mm} + \Re(\mathbf{y}_m) - \sum_{j \neq m}^{N_r} \lambda_j \Re(\mathbf{K})_{mj} - \epsilon(sgn(\tilde{\lambda}_m^{new}))$$

where $\Re(y_m) - \sum_{j \neq m}^{N_r} \lambda_j \Re(\mathbf{K})_{mj} = \Re(y_m) - \sum_{j=1}^{N_r} \lambda_j \Re(\mathbf{K}_{mj}) = \Phi_m^r + \lambda_m \Re(\mathbf{K}_{mm})$ and λ_m

denotes the mth optimization variable before update. Therefore (74) can be rewritten as

$$\frac{\partial \theta_m^r}{\partial \tilde{\lambda}_m^{new}} = -\tilde{\lambda}_m^{new}\Re(\mathbf{K})_{mm} + \Phi_m^r + \lambda_m \Re(\mathbf{K})_{mm} - \epsilon(sgn(\tilde{\lambda}_m^{new})) \tag{74}$$

$$\Rightarrow \tilde{\lambda}_m^{new} = \lambda_m + \frac{\Phi_m^r - \epsilon(sgn(\tilde{\lambda}_m^{new}))}{\Re(\mathbf{K})_{mm}}, \tag{75}$$

The update of $\tilde{\lambda}_m^{new}$ is completed by clipping based on (69)

$$\lambda_m^{new} = [\tilde{\lambda}_m^{new}]_{-CR'(\xi_m)}^{CR'(\xi_m)} \tag{76}$$

where $[]_a^b$ denotes clipping function

$$[x]_a^b = \begin{cases} a \quad if \quad x \leq a \\ x \quad if \quad a < x < b \\ b \quad if \quad x \geq b \end{cases} \tag{77}$$

based on (72), The maximal gain of the sub dual objective function with respect to the mth optimization variable is

$$\begin{aligned} &\bigtriangledown \theta_m^r = \theta_m^r(\lambda_m + \sigma_m) - \theta_m^r(\lambda_m) \\ &= -\frac{1}{2}\sigma_m^2 \Re(\mathbf{K})_{mm} + \Phi_m^r \sigma_m - \epsilon(|\lambda_m^{new}| - |\lambda_m|) \\ &= \sigma_m[-\frac{1}{2}\sigma_m \Re(\mathbf{K})_{mm} + \Phi_m^r] - \epsilon(|\lambda_m^{new}| - |\lambda_m|), \end{aligned} \tag{78}$$

In 1-D searching procedure, the optimization variable whose update can achieve the maximum gain of sub dual objective function is chosen, assume the kth optimization variable is chosen, based on the this principle

$$k = \arg \max_{(m=1,2,\ldots,N_r)} \bigtriangledown \theta_m^r. \tag{79}$$

Notice in (74), $sgn(\tilde{\lambda}_m^{new})$ is unknown before λ_m is updated, therefore two conditions $sgn(\tilde{\lambda}_m^{new}) = 1$ and $sgn(\tilde{\lambda}_m^{new}) = -1$ are considered. The final decision of $sgn(\tilde{\lambda}_m^{new})$ is made by comparing the corresponding gains of the sub dual objective functions in (78).

B. Double Direction (2-D) Solver

Although the omission of offset in the CSVR-MIMO detector makes 1-D solver possible, however, recent work in [36] 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 optimal 2-D solver uses the same principle as 1-D solver, $|S| = 2$, assume the work set consists of the mth and nth optimization variables, that are $\Lambda_S = [\lambda_m, \lambda_n]$. Based on (72), the sub dual objective function can be written as

$$\begin{aligned}
&\max_{\lambda_m,\lambda_n} \quad \theta_{m,n}^r = -\frac{1}{2}[(\tilde{\lambda}_m^{new})^2 \Re(\mathbf{K})_{mm} + (\tilde{\lambda}_n^{new})^2 \Re(\mathbf{K})_{nn} + 2\tilde{\lambda}_m^{new}\tilde{\lambda}_n^{new} \Re(\mathbf{K})_{mn}] - \\
&\tilde{\lambda}_m^{new} \sum_{j \neq m,n}^{N_r} \lambda_j \Re(\mathbf{K})_{mj} - \tilde{\lambda}_n^{new} \sum_{j \neq m,n}^{N_r} \lambda_j \Re(\mathbf{K})_{nj} + \Re(\mathbf{y}_m)\tilde{\lambda}_1^{new} + \Re(\mathbf{y}_n)\tilde{\lambda}_n^2 \\
&-\epsilon(|\lambda_m^{new}| + |\tilde{\lambda}_n^{new}|),
\end{aligned} \tag{80}$$

Based on (80), the partial derivatives of $\theta^r_{m,n}$ with respect to $\tilde{\lambda}^{new}_m$ and $\tilde{\lambda}^{new}_n$ are

$$\frac{\partial \theta^r_{m,n}}{\partial \tilde{\lambda}^{new}_m} = -\tilde{\lambda}^{new}_m \Re(\mathbf{K})_{mm} - \tilde{\lambda}^{new}_n \Re(\mathbf{K})_{mn} - \sum_{j \neq m,n}^{N_r} \lambda_j \Re(\mathbf{K})_{mj} + \Re(\mathbf{y}_m) - \epsilon sgn(\tilde{\lambda}^{new}_m) =$$
$$-\tilde{\lambda}^{new}_m \Re(\mathbf{K})_{mm} - \tilde{\lambda}^{new}_n \Re(\mathbf{K})_{mn} + \Phi^r_m + \lambda_m \Re\mathbf{K}_{mm} + \lambda_n \Re(\mathbf{K})_{mn} - \epsilon sgn(\tilde{\lambda}^{new}_m) = 0 \qquad (81)$$
$$\frac{\partial \theta^r_{m,n}}{\partial \tilde{\lambda}^{new}_n} = -\tilde{\lambda}^{new}_n \Re(\mathbf{K})_{nn} - \tilde{\lambda}^{new}_m \Re(\mathbf{K})_{mn} - \sum_{j \neq m,n}^{N_r} \lambda_j \Re(\mathbf{K})_{nj} + \Re(\mathbf{y}_n) - \epsilon sgn(\tilde{\lambda}^{new}_n) =$$
$$-\tilde{\lambda}^{new}_n \Re(\mathbf{K})_{nn} - \tilde{\lambda}^{new}_m \Re(\mathbf{K})_{mn} + \Phi^r_n + \lambda_m \Re(\mathbf{K})_{mn} + \lambda_n \Re(\mathbf{K})_{nn} - \epsilon sgn(\tilde{\lambda}^{new}_n) = 0 \qquad (82)$$

where $\frac{\partial |x|}{\partial x} = sgn(x)$ denotes the sign of x. Based on (81) and (82) we have

$$(\tilde{\lambda}^{new}_m - \lambda_n)\Re(\mathbf{K})_{mm} = \Phi^r_m - \epsilon sgn(\tilde{\lambda}^{new}_m) - (\tilde{\lambda}^{new}_n - \lambda_n)\Re(\mathbf{K})_{mn} \qquad (83)$$

$$(\tilde{\lambda}^{new}_n - \lambda_n)\Re(\mathbf{K})_{nn} = \Phi^r_n - \epsilon sgn(\tilde{\lambda}^{new}_n) - (\tilde{\lambda}^{new}_m - \lambda_m)\Re(\mathbf{K})_{mn} \qquad (84)$$

hence based on (83) and (84), the update formulas of $\tilde{\lambda}^{new}_m$ and $\tilde{\lambda}^{new}_n$ are given by

$$\tilde{\lambda}^{new}_m = \lambda_m + \frac{\Phi^r_m \Re(\mathbf{K})_{nn} - \Phi^r_n \Re(\mathbf{K})_{mn} - \epsilon[sgn(\tilde{\lambda}^{new}_m)\Re(\mathbf{K})_{nn} - sgn(\tilde{\lambda}^{new}_n)\Re(\mathbf{K})_{mn}]}{\Re(\mathbf{K})_{mm}\Re(\mathbf{K})_{nn} - (\Re(\mathbf{K})_{mn})^2} \qquad (85)$$

$$\tilde{\lambda}^{new}_n = \lambda_n + \frac{\Phi^r_n \Re(\mathbf{K})_{mm} - \Phi^r_m \Re(\mathbf{K})_{mn} - \epsilon[sgn(\tilde{\lambda}^{new}_n)\Re(\mathbf{K})_{mm} - sgn(\tilde{\lambda}^{new}_m)\Re(\mathbf{K})_{mn}]}{\Re(\mathbf{K})_{mm}\Re(\mathbf{K})_{nn} - (\Re(\mathbf{K})_{mn})^2} \qquad (86)$$

Then the updated optimization variables are clipped by constraint based on (69)

$$\lambda_i^{new} = [\tilde{\lambda}_i^{new}]_{-CR'(\xi_i)}^{CR'(\xi_i)},$$
$$i = m, n \tag{87}$$

Based on (64), the gain of the sub dual objective function can be written as

$$\bigtriangledown\theta_{mn}^r = -\frac{1}{2}[\sigma_m^2\Re(\mathbf{K})_{mm} + \sigma_n^2\Re(\mathbf{K})_{nn} + 2\sigma_m\sigma_n\Re(\mathbf{K})_{mn}] + \Phi_m^r\sigma_m + \Phi_n^r\sigma_n$$
$$-\epsilon(|\lambda_m^{new}| - |\lambda_m| + |\lambda_n^{new}| - |\lambda_n|), \tag{88}$$

assume the ith and jth optimization variables are chosen, the optimization variables in 2-D solver have the same update rule as that of 1-D solver, that is

$$[i, j] = \arg\max_{m,n=1,2,\ldots,N_r} \bigtriangledown\theta_{m,n}^r. \tag{89}$$

Similar to single direction solver, notice in (85) and (86), $sgn(\tilde{\lambda}_m^{new})$ and $sgn(\tilde{\lambda}_n^{new})$ are unknown before λ_m and λ_n are updated, therefore four conditions are considered

$$sgn(\tilde{\lambda}_m^{new}) = 1, sgn(\tilde{\lambda}_n^{new}) = 1$$
$$sgn(\tilde{\lambda}_m^{new}) = 1, sgn(\tilde{\lambda}_n^{new}) = -1$$
$$sgn(\tilde{\lambda}_m^{new}) = -1, sgn(\tilde{\lambda}_n^{new}) = 1$$
$$sgn(\tilde{\lambda}_m^{new}) = -1, sgn(\tilde{\lambda}_n^{new}) = -1$$

The final decisions of $sgn(\tilde{\lambda}_m^{new})$ and $sgn(\tilde{\lambda}_n^{new})$ are made by comparing the corresponding possible gains of the sub objective functions in (88).

Modify (88), we have

$$\bigtriangledown\theta_{mn}^{r} = \sigma_m[-\frac{1}{2}\sigma_m\Re(\mathbf{K})_{mm} + \Phi_m^r] - \epsilon(|\lambda_m^{new}| - |\lambda_m|) + \sigma_n[-\frac{1}{2}\sigma_n\Re(\mathbf{K})_{nn} + \Phi_n^r] - \epsilon(|\lambda_n^{new}| - |\lambda_n|)$$
$$-\sigma_m\sigma_n\Re(\mathbf{K})_{mn}, \tag{90}$$

then recall the gain of the sub dual objective function of 1-D solver in (78), we obtain

$$\bigtriangledown\theta_{ij}^{r} = \bigtriangledown\theta_i^r + \bigtriangledown\theta_j^r - \sigma_i\sigma_j\Re(\mathbf{K})_{ij}, \tag{91}$$

where $\bigtriangledown\theta_i^r$, $\bigtriangledown\theta_j^r$ denote gains of the sub dual objective functions of 1-D solver with ith and jth optimization variables are updated.

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

From (88), we can observe that the gain of the sub dual objective function of 2-D solver is a summation of the gains of the sub dual objective functions of two independent 1-D solver and a correlation term $\sigma_i\sigma_j\Re(\mathbf{K})_{ij}$.

The optimal 2-D work set $[i, j]$ can be determined by brute-force searching manner, which requires $O(n^2)$ searching times. Based on (88), we can approximate optimal 2-D searching strategy by the combinations of optimal 1-D searching approach, we will prove in the large MIMO systems, when N_t is sufficient large, this approximation is very effective. Here we propose two types of combined 1-D searching strategy:

1) One-shot 1-D combined Searching: do one round 1-D searching and calculate all the 1-D gains based on (78), then choose the two optimization variables indexed by i and j, whose update can achieve the first and the second largest gain of the sub dual objective functions.

2) Sequential 1-D combined Searching: do two rounds 1-D searchings, in the first round find optimization variable indexed by i whose update can achieve the maximal 1-D gain of the sub dual objective functions based on (78), then update the ith optimization variable. In the second round, find jth optimization variable whose update can achieve the maximal 1-D gain of the sub dual objective functions.

The effectiveness of 1-D approximation strategies are majorly determined by the ratio $\frac{\sigma_i \sigma_j \Re(\mathbf{K})_{ij}}{\bigtriangledown\theta_i^r + \bigtriangledown\theta_j^r}$. Hence we provide theoretical analyses based on the view of channel hardening phenomenon. It can be proved that in the large MIMO systems, the correlation term $\sigma_i \sigma_j \Re(\mathbf{K}_{ij})$is ignorable comparing to 1-D gains of the sub objective function. Prior to the theoretical analyse, we first investigate some mathematical properties of channel hardening (to be completed).

V. Stopping Criteria

As discussed in sectionII-C, in dual objective function, the complementary KKT conditions are the measurement of the proximity of current solution to the optimal solution. The feasibility gap, which is a effective stopping criteria used in SVR and SVM, can implicitly monitor whether the KKT complementary conditions are satisfied. The feasibility gap measures the gap between the values of primal objective function and dual objective function. To elaborate a little further, recall the saddle point conditions in sectionII-C

$$L(\mathbf{w}, \bar{\mathbf{a}}) \geq L(\bar{\mathbf{w}}, \bar{\mathbf{a}}) \geq L(\bar{\mathbf{w}}, \mathbf{a}) \tag{92}$$

Where $(\bar{\mathbf{w}}, \bar{\mathbf{a}})$ are the solution to primal objective function, the KKT complementary conditions are satisfied, that is

$$\bar{\alpha}_i c_i(\bar{\mathbf{w}}) = 0, \tag{93}$$

we have

$$L(\bar{\mathbf{w}}, \bar{\mathbf{a}}) = f(\bar{\mathbf{w}}) + \sum_{i=1}^{N_r} \bar{\alpha}_i c_i(\bar{\mathbf{w}}) = f(\bar{\mathbf{w}}) \tag{94}$$

Let $(\tilde{\mathbf{w}}, \tilde{\mathbf{a}})$ denote the variable pair that satisfies the first and second KKT conditions in (17) and (18)

$$\partial_{\mathbf{w}} L(\tilde{\mathbf{w}}, \tilde{\mathbf{a}}) = \partial_{\mathbf{w}} f(\tilde{\mathbf{w}}) + \sum_{i=1}^{L} \tilde{a}_i \partial_{\mathbf{w}} c_i(\tilde{\mathbf{w}}) = 0, \tag{95}$$

$$c_i(\tilde{\mathbf{w}}) \leq 0, i = 1, 2, \ldots L \tag{96}$$

based on (95), we have

$$L(\tilde{\mathbf{w}}, \tilde{\mathbf{a}}) = f(\tilde{\mathbf{w}}) + \sum_{i=1}^{L} \tilde{\alpha}_i c_i(\tilde{\mathbf{w}}) \leq f(\tilde{\mathbf{w}}), \tag{97}$$

based on Theorem 6.27 [26], the following equalities hold

$$f(\tilde{\mathbf{w}}) \geq f(\bar{\mathbf{w}}) = L(\bar{\mathbf{w}}, \bar{\mathbf{a}}) \geq L(\tilde{\mathbf{w}}, \tilde{\mathbf{a}}) = f(\tilde{\mathbf{w}}) + \sum_{i=1}^{L} \tilde{\alpha}_i c_i(\tilde{\mathbf{w}}) \tag{98}$$

Therefore the feasibility gap is defined by the gap between the primal objective function $f(\tilde{\mathbf{w}})$ and the dual objective function $L(\tilde{\mathbf{w}}, \tilde{\mathbf{a}})$

$$f(\tilde{\mathbf{w}}) - L(\tilde{\mathbf{w}}, \tilde{\mathbf{a}}) = -\sum_{i=1}^{L} \tilde{\alpha}_i c_i(\tilde{\mathbf{w}}) \geq 0 \tag{99}$$

If the feasibility gap is vanished, based on (98), we have $f(\bar{\mathbf{w}}) = f(\tilde{\mathbf{w}})$, thus $\tilde{\mathbf{w}} = \bar{\mathbf{w}}$, $\tilde{\mathbf{a}} = \bar{\mathbf{a}}$. Furthermore, because $f(\tilde{\mathbf{w}}) - L(\tilde{\mathbf{w}}, \tilde{\mathbf{a}}) = -\sum_{i=1}^{L} \tilde{\alpha}_i c_i(\tilde{\mathbf{w}})$, the vanish of feasibility gap indicates $\sum_{i=1}^{L} \tilde{\mathbf{a}} c_i(\tilde{\mathbf{w}}) = 0$. Because $\tilde{\alpha}_i c_i(\tilde{\mathbf{w}}) \leq 0, i = 1, 2, \ldots, L$, we have $\tilde{\alpha}_i c_i(\tilde{\mathbf{w}}) = 0, i = 1, 2, \ldots, L$, the KKT complementary conditions are satisfied.

In CSVR detector for the large MIMO systems, the primal objective function is

$$f(\tilde{\mathbf{w}}, \xi_i^r, \hat{\xi}_i^r, \xi_i^i, \hat{\xi}_i^i) = \frac{1}{2}||\tilde{\mathbf{w}}||_{\mathbb{H}}^2 + C\sum_{i=1}^{N_r}[R(\xi_i^r) + R(\hat{\xi}_i^r) + R(\xi_i^i) + R(\hat{\xi}_i^i)], \tag{100}$$

where $\tilde{\mathbf{w}}$ satisfies the KKT conditions in (95) and (96). Define $\lambda^r = \alpha - \hat{\alpha}$, $|\lambda^r| = \alpha + \hat{\alpha}$ and $\lambda^i = \beta - \hat{\beta}$, $|\lambda^i| = \beta + \hat{\beta}$, $\Lambda^r = [\lambda_1^r, \lambda_2^r, \ldots, \lambda_L^r]^T$ and $\Lambda^i = [\lambda_1^i, \lambda_2^i, \ldots, \lambda_L^i]^T$, (45) can be rewritten as

$$||\tilde{\mathbf{w}}||_{\mathbb{H}}^2 =< (\Lambda^r)^T, \Re(\mathbf{K})\Lambda^r > + < (\Lambda^i)^T, \Re(\mathbf{K})\Lambda^i > -2 < \Lambda^r, \Im(\mathbf{K})\Lambda^i >, \tag{101}$$

Therefore we obtain

$$f(\tilde{\mathbf{w}}, \xi_i^r, \hat{\xi}_i^r, \xi_i^i, \hat{\xi}_i^i) = \frac{1}{2} < (\Lambda^r)^T, \Re(\mathbf{K})\Lambda^r > + \frac{1}{2} < (\Lambda^i)^T, \Re(\mathbf{K})\Lambda^i > - < \Lambda^r, \Im(\mathbf{K})\Lambda^i > + C\sum_{i=1}^{N_r}[R(\xi_i^r) + R(\hat{\xi}_i^r) + R(\xi_i^i) + R(\hat{\xi}_i^i)], \tag{102}$$

Based on (51), the dual objective function can be rewritten as

$$\theta(\Lambda^r, \Lambda^i) = -\frac{1}{2} < (\Lambda^r)^T, \Re(\mathbf{K})\Lambda^r > -\frac{1}{2} < (\Lambda^i)^T, \Re(\mathbf{K})\Lambda^i > + < \Re(\mathbf{y})^T, \Lambda^r > + < \Im(\mathbf{y})^T, \Lambda^i >$$
$$-\epsilon < \mathbf{e}^T, (|\Lambda^r| + |\Lambda^i|) > +C \sum_{i=1}^{N_r} [\tilde{R}(\xi_i^r) + \tilde{R}(\hat{\xi}_i^r) + \tilde{R}(\xi_i^i) + \tilde{R}(\hat{\xi}_i^i)], \tag{103}$$

Based on (102) and (103), the feasibility gap is obtained

$$G(\lambda^r, \lambda^i) = f(\tilde{\mathbf{w}}, \xi_i^r, \hat{\xi}_i^r, \xi_i^i, \hat{\xi}_i^i) - \theta(\Lambda^r, \Lambda^i) =< (\Lambda^r)^T, \Re(\mathbf{K})\Lambda^r > + < (\Lambda^i)^T, \Re(\mathbf{K})\Lambda^i > - < \Re(\mathbf{y}^T), \Lambda^r >$$
$$- < \Im(\mathbf{y})^T, \Lambda^i > +\epsilon < \mathbf{e}^T, (|\Lambda^r| + |\Lambda^i|) > +C \sum_{i=1}^{N_r} [\xi_i^r R'(\xi_i^r) + \hat{\xi}_i^r R'(\hat{\xi}_i^r) + \xi_i^i R'(\xi_i^i) + \hat{\xi}_i^i R'(\hat{\xi}_i^i)] -$$
$$< \Lambda^r, \Im(\mathbf{K})\Lambda^i > . \tag{104}$$

As we explained in section II-B, the choice of risk function is determined by distribution of noise, as to Gaussian noise, the risk function is

$$R(u) = \frac{1}{2}u^2, \tag{105}$$

hence

$$\tilde{R}(u) = R(u) - uR'(u) = -\frac{1}{2}u^2, \tag{106}$$

In ϵ-SVR, the objective to exploit slack variables is to compensate the influences from the outliers that exceed the ϵ-tube which are caused by the noise. Therefore in ϵ-SVR, ξ_i^r and $\hat{\xi}^r$ are defined as

$$\xi_i^r = \max(0, \Re(\mathbf{y}_i) - \Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - \epsilon) \tag{107}$$

$$\hat{\xi}_i^r = \max(0, \Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - \Re(\mathbf{y}_i) - \epsilon) \tag{108}$$

Because the distance between the estimations $\Re(<\mathbf{h}_i,\mathbf{w}>_{\mathbb{H}})$ and the observations $\Re(\mathbf{y}_i)$ can only exceeds the ϵ-tube in one direction, therefore there is at most one of ξ_i^r and $\hat{\xi}_i^r$ can be non zero that is $\xi_i^r\hat{\xi}_i^r = 0$. Therefore the risk function can be rewritten as

$$R(\xi_i^r) + R(\hat{\xi}_i^r) = \frac{1}{2}|\Re(\mathbf{y}_i) - \Re(<\mathbf{h}_i,\mathbf{w}>_{\mathbb{H}})|_\epsilon^2 \tag{109}$$

$$R(\xi_i^i) + R(\hat{\xi}_i^i) = \frac{1}{2}|\Im(\mathbf{y}_i) - \Im(<\mathbf{h}_i,\mathbf{w}>_{\mathbb{H}})|_\epsilon^2 \tag{110}$$

where $|\cdot|_\epsilon$ denotes ϵ insensitive function as mentioned in section II-B. Recall (44),

$$<\mathbf{h}_i,\mathbf{w}>_{\mathbb{H}}= \sum_{j=1}^{N_r}\lambda_j^r\Re(\mathbf{K})_{ij} - \sum_{j=1}^{N_r}\lambda_j^i\Im(\mathbf{K})_{ij} + i(\sum_{j=1}^{N_r}\lambda_j^r\Im(\mathbf{K})_{ij} + \sum_{j=1}^{N_r}\lambda_j^i\Re(\mathbf{K})_{ij}), \tag{111}$$

we obtain

$$\Re(\mathbf{y}_i) - \Re(<\mathbf{h}_i,\mathbf{W}>_{\mathbb{H}}) = \Re(\mathbf{y}_i) - \sum_{j=1}^{N_r}\lambda_j^r\Re(\mathbf{K})_{ij} + \sum_{j=1}^{N_r}\lambda_j^i\Im(\mathbf{K})_{ij} \tag{112}$$

$$\Im(\mathbf{y}_i) - \Im(<\mathbf{h}_i,\mathbf{w}>_{\mathbb{H}}) = \Im(\mathbf{y}_i) - \sum_{j=1}^{N_r}\lambda_j^i\Re(\mathbf{K})_{ij} - \sum_{j=1}^{N_r}\lambda_j^r\Im(\mathbf{K})_{ij} \tag{113}$$

Two intermediate variables Φ and Ψ are defined

$$\Phi^r = \Re(\mathbf{y}) - \Re(\mathbf{K})\lambda^r; \Phi^i = \Im(\mathbf{y}) - \Re(\mathbf{K})\lambda^i \tag{114}$$

$$\Psi^r = \Im(\mathbf{K})\lambda^i; \Psi^i = -\Im(\mathbf{K})\lambda^r \tag{115}$$

therefore (112) and (113) can be rewritten as

$$\Re(\mathbf{y}_i) - \Re(< \mathbf{h}_i, \mathbf{W} >_{\mathbb{H}}) = \Phi_i^r + \Psi_i^r \tag{116}$$

$$\Im(\mathbf{y}_i) - \Im(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) = \Phi_i^i + \Psi_i^i \tag{117}$$

Therefore based on (116) and (117), (109) and (110) can be rewritten as

$$R(\xi_i^r) + R(\hat{\xi}_i^r) = \frac{1}{2}|\Phi_i^r + \Psi_i^r|_\epsilon^2 \tag{118}$$

$$R(\xi_i^i) + R(\hat{\xi}_i^i) = \frac{1}{2}|\Phi_i^i + \Psi_i^i|_\epsilon^2 \tag{119}$$

Thus based on (118) and (119), the feasibility gap in (104) can be rewritten as

$$G(\Lambda^r, \Lambda^i) =< (\Lambda^r)^T, \Re(\mathbf{K})\Lambda^r > + < (\Lambda^i)^T, \Re(\mathbf{K})\Lambda^i > - < \Re(\mathbf{y})^T, \Lambda^r > - < \Im(\mathbf{y})^T, \Lambda^i >$$

$$+\epsilon < \mathbf{e}^T, (|\Lambda^r| + |\Lambda^i|) > +C \sum_{i=1}^{N_r} [(|\Phi_i^r + \Psi_i^r|)_\epsilon^2 + (|\Phi_i^i + \Psi_i^i|)_\epsilon^2] - < \Lambda^r, \Im(\mathbf{K})\Lambda^i > . \tag{120}$$

Based on dual objective function in (103), (120) can be rewritten as

$$G(\Lambda^r, \Lambda^i) =< \Re(\mathbf{y})^T, \Lambda^r > + < \Im(\mathbf{y})^T, \Lambda^i > -\epsilon < \mathbf{e}^T, (|\Lambda^r| + |\Lambda^i|) > -2\theta(\Lambda_i, \Lambda_j)$$

$$- < \Lambda^r, \Im(\mathbf{K})\Lambda^i > . \tag{121}$$

The following ratio of the feasibility gap $G(\Lambda^r, \Lambda^i)$ is used to measure the proximity of current solution to the optimal solution. The algorithm stops when the ratio satisfies a certain tolerance.

$$\frac{G}{G+\theta} \tag{122}$$

A. Update Φ, Ψ and G

Φ, Ψ and G are updated in each iteration along with the optimization variables updated. The pseudo code to update Φ, Ψ and G is given.

Based on the definitions of Φ and Ψ in (114) and (115), we have the following procedure to update Φ and Ψ in real channel and imaginary channel, assume the optimization variables updated in each channel are indexed by 1 and 2, let $\sigma_i^r = (\lambda_i^r)^{new} - \lambda_i^r$ and $\sigma_i^i = (\lambda_i^i)^{new} - \lambda_i^i$ denote the differences between the updated optimization variables and the old optimization variables.

---

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 cost function term in (120) are updated by **Procedure 3** and **Procedure 4**. The cost function terms are denoted by χ^r in real channel and χ^i in imaginary channel.

$$\chi^r = \sum_i^{N_r} |\Phi_i^r + \Psi_i^r|_\epsilon^2 \tag{123}$$

$$\chi^i = \sum_i^{N_r} |\Phi_i^i + \Psi_i^i|_\epsilon^2 \tag{124}$$

procedure 3. UPDATE COST 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 COST 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 (121) is shown in **Procedure 5**. Assume the indexes of optimization variables updated in real channel is i and j, the indexes of optimization variables updated in imaginary channel is m and f. Define the work sets in real and imaginary channels $S^r = [i, j]$ and $S^i = [m, f]$. Define $(\chi^r)^{new}$ and $(\chi^i)^{new}$ are the updated cost function terms based on **Procedure 3** and **Procedure 4**, the update process for the term $\theta(\Lambda^r, \Lambda^i)$ in (121) can be written as

$$\bigtriangledown\theta(\Lambda^r_{S^r}, \Lambda^i_{S^i}) = \bigtriangledown\theta^r_{S^r} + \bigtriangledown\theta^i_{S^i} - \frac{C}{2}((\chi^r)^{new} + (\chi^i)^{new} - \chi^r - \chi^i), \tag{125}$$

Where θ_S^r is the gain of the sub dual objective function as shown in (72),

$$\bigtriangledown\theta_{S^r}^r = -\frac{1}{2}(\Sigma_{S^r}^r)^T\Re(\mathbf{K})_{S^rS^r}\Sigma_{S^r}^r + (\Phi_{S^r}^r)^T\Sigma_{S^r}^r - \epsilon < \mathbf{e}_{S^r}^T, |(\Lambda_{S^r}^r)^{new}| - |\Lambda_{S^r}^r| > \quad (126)$$

$$\bigtriangledown\theta_{S^i}^i = -\frac{1}{2}(\Sigma_{S^i}^i)^T\Re(\mathbf{K})_{S^iS^i}\Sigma_{S^i}^i + (\Phi_{S^i}^i)^T\Sigma_{S^i}^i - \epsilon < \mathbf{e}_{S^i}^T, |(\Lambda_{S^i}^i)^{new}| - |\Lambda_{S^i}^i| > \quad (127)$$

---

procedure 5. UPDATE G
- $G+ = Re(\mathbf{y}_1)\sigma_i^r + Re(\mathbf{y}_2)\sigma_j^r$
- $G+ = Im(\mathbf{y}_1)\sigma_m^i + Re(\mathbf{y}_2)\sigma_f^i$
- $G- = \epsilon(|\lambda_i^r + \sigma_i^r| - |\lambda_i^r| + |\lambda_j^r + \sigma_j^r| - |\lambda_j^r|)$
- $G- = \epsilon(|\lambda_m^i + \sigma_m^i| - |\lambda_m^i| + |\lambda_f^i + \sigma_f^i| - |\lambda_f^i|)$
- $G- = 2(\bigtriangledown\theta(\Lambda_{S^r}^r, \Lambda_{S^i}^i))$ ▷ Update sub objective function based on (125)
- $G- = (\Sigma_{S^r}^r)^T\Im(\mathbf{K}_{S^rS^i})\Sigma_{S^i}^i + (\Sigma_{S^r}^r)^T\Psi_{S^r}^r + (\Sigma_{S^i}^i)^T\Psi_{S^i}^i$ ▷ Update $< \Lambda^r, \Im(\mathbf{K})\Lambda^i >$

end procedure

Pseudo code for two types of combined 1-D searching solver are given by **Procedure 6** and **Procedure 7**.

---

procedure 6. ONE-SHOT 1-D COMBINED SEARCHING SOLVER

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 IV-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 IV-B

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

end procedure

---

procedure 7. SEQUENTIAL 1-D COMBINED SEARCHING SOLVER

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 IV-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 IV-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 $\lambda_{1st_1}^r(\lambda_{1st_1}^i)$ and $\lambda_{1st_2}^r(\lambda_{1st_2}^i)$ ▷ Based on double direction solver IV-B

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

end procedure

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

VI. Computer Simulations

Computer simulations are 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.

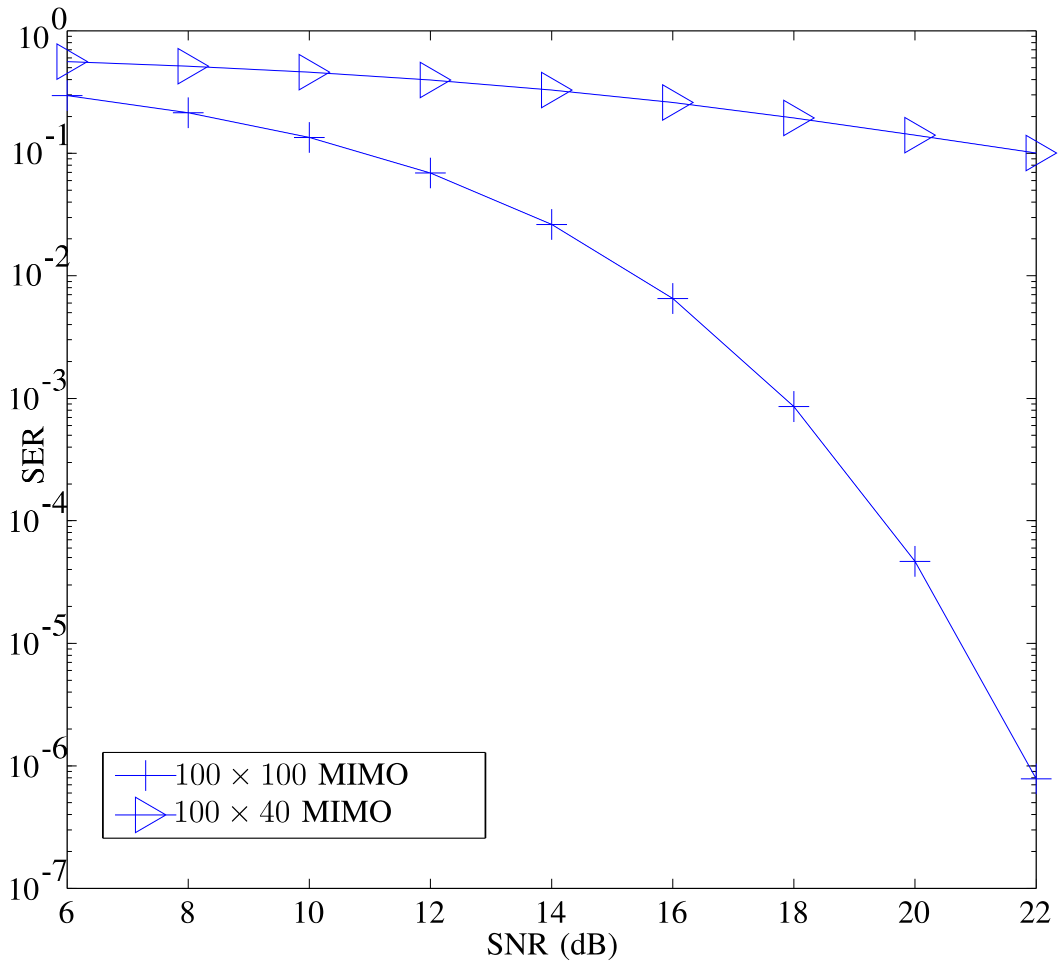

Fig. 3. SER performance of 100×100 and 100×40 MIMO system

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(**y**,**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.
  call combined 1-D searching solver in **Procedure 6** and **Procedure 7** ▷ find and update two optimization variables in the real and imaginary channel
  Step 4. **Procedure 5** update $\frac{G}{G+\theta}$, go back to step 2.
  Step 5.
  $\tilde{\mathbf{x}} = \mathbf{H}^H(\Lambda^r + i\Lambda^i)$ ▷ reconstruct $\mathbf{x}$
  $\mathbf{x} = \mathbb{Q}(\tilde{x})$ ▷ $\mathbb{Q}(\cdot)$ denotes quantization operation based on symbol constellation
  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.