Report

Tianpei Chen

Department of Electrical and Computer Engineering

McGill University

September 8, 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) and the limited radio frequency spectrum over a complex space-time varying environment. As the most promising technology for solving this problem, Multiple Input Multiple Output (MIMO) technology has been of immense research interest over the last several tens of years and become mature, which is incorporated into the emerging wireless broadband standard like 802.11ac [1] long-term evolution (LTE) [2]. The core idea of MIMO system is to use multiple antennas at both transmitting and receiving end, so that multiplexing gain (multiple parallel spatial data pipelines that can improve bandwidth efficiency) and diversity gain (better reliability of communication link) is obtained by employing spatial domain. Large MIMO (also called Massive MIMO) is an upgrade version of conventional MIMO system, it equips unprecedentedly hundreds of low power low price antennas at base station (BS), serving several tens of terminals simultaneously, It can achieve full potential of conventional MIMO system while providing additional power efficiency as well as system robustness [3] [4].

The price paid for large MIMO system is the increasing complexities for signal processing at both transmitting and receiving end. Uplink Detector is one of the key components in large MIMO system. With orders magnitude more antennas equipped at BS, benefit and challenge coexist in designing of detection algorithms for large MIMO uplink, on the one hand, large number of receive antennas provide potential of large diversity gain, on the other hand, complexity of the algorithm becomes extremely crucial to make system practical.

The optimal maximum likelihood detector (MLD) for MIMO system requires the complexity increase exponentially with number of transmitted antennas with a factor of the size of size of constellation, which is prohibitive in practical implementations. Sphere Decoder (SD) [5] is the most prominent algorithm that utilizes lattice structure of MIMO system. Its variant fixed complexity sphere decoder (FCSD) [6] make it possible to achieve near optimal performance with a fixed complexity under different signal to noise ratio (SNR). However, all the algorithms that based on lattice structure have the same shortage - their complexities increases exponentially with a factor of the size of symbol constellation. Therefore, they are prohibitive when it comes to a high order modulation scheme, for example in IEEE 802.11ac standard [1], the modulation scheme is 256QAM.

Suboptimal linear detectors (LD) like minimum mean square error (MMSE) and zero forcing (ZF) along with their sequential interference cancellation with optimized ordering (OSIC) counterparts [7] [8] [9], which have good performance for low loading factor in massive MIMO system (that is the number of receive antennas is much larger than the number of transmit antennas) [10]. In the last several years, a set of detection algorithms are proposed with complexities that is comparable with LD-OSIC and suboptimal performance can be achieved. The local search algorithm, such as likelihood ascend searching (LAS) [11] [12], an theoretical analysis of upper bound of bit error rate (BER) and lower bound of on asymptotic multiuser efficiency for the LAS detector was presented [13]. Layered Tabu search algorithm presented in [14] is superior to the LAS algorithms because it can move away to new searching area to avoid local minimal. Message passing detectors based on belief propagation (BF) and Gaussian Approximation (GA) [15] [16] [17] [18]. Markov Chain Monte Carlo (MCMC) algorithm [19] and Lattice Reduction (LR) aided detectors [20].

Firmly grounded in framework of statistical learning theory, Support Vector Machine (SVM) has become a powerful tool to solve real world supervised learning problems such as classification, regression and prediction. SVM method is a nonlinear generalization of Generalized Portrait algorithm developed by Vapnik in 1960s [21] [22], which can give good generalization performance to unseen data [23].

Research and industry interest of SVM boosted since 1990s, promoted by related works of Vapnik and co-workers at AT& T Bell laboratory [24] [25] [26] [27] [28] [29] Moreover, the kernel based methods [23] carries out nonlinear learning task by mapping input data sets into high dimensional feature space, then replacing inner product of feature mappings by computational inexpensive kernel functions discarding the actual structure of the feature space. This rational is supported mathematically by Reproducing Kernel Hilbert Space (RKHS). Based on the same regularized risk function principle, -Support Vector Regression (ϵ-SVR) was developed [26] [30].

Like SVM, ϵ-SVR solving original optimization problem by transforming it into Lagrange dual optimization problem, which can be solved by Quadratic Programming (QP), Sequential Minimal Optimization (SMO) algorithm was proposed as a fast algorithm to solve this QP problem by decompose the it into sub QP problems and solve them analytically [31], therefore, the computational intensive numerical method can be avoided. A more general method is decomposition solver, which refers to a set of algorithms that separate the optimization variables (Lagrange multipliers) into two sets W and N, W is the work set and N contains the remaining optimization variables. In each iteration, only the optimization variables in work set is optimized while keeping other variables fixed. SMO algorithm is an extreme case of decomposition solver. An important issue of decomposition solver is the choice of work set, one strategy is to choose KarushKuhnTucker (KKT) condition violators, and final converge can be guaranteed [32]. SMO algorithm restricts the size of work set to 2, because of linear constraint in dual problem that inducted by offset. In [33], a method to train SVM without offset was proposed, with the comparable performance to the SVM with offset. The authors work demonstrates that with the combination of two single optimization variable work set selection strategies which requires searching time $O(n)$ and update a work set size of two in each iteration, this method can achieve a iteration time as few as that searching over all pairs of optimization variables which requires $O(n^2)$ searching times.

Until now, although the mathematical foundation of kernel based methods is RKHS which is defined in complex domain, the most of practitioners are dealing with real data set. 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 [34], which can process the complex data set purely in complex domain. The simulation of channel realization and equalization in [34] demonstrate a better performance as well as reduced complexity comparing to simply split learning task into two real case by real kernel. Based on this work, we construct a prototype of a complexity-performance controllable detector for large MIMO based on dual channel complex SVR. The detector can work in two parallel real SVR pipeline which can be solved independently. Moreover, only real part of kernel matrix is needed in both channel. This means a large amount of computation can be reduced. Based on the discrete time MIMO channel model, In our regression model, this CSVR-detector is constructed without offset, Therefore, for each real SVR without offset, in principle, only one variable is needed to be updated in each iteration, In our prototype, we propose a sequential single optimization variable searching strategy that find two optimization variable sequentially, which can approximate the optimal double optimization variables searching strategy.

II. System Model

Consider a large MIMO uplink multiplexing system with N_t users, each user has one transmit antenna. The number of receive antennas at Base Station (BS) is N_r, $N_r \geq N_t$. Typically large MIMO systems have hundreds of antennas at BS serving several tens of terminals, as shown in Fig 1.

Uncoded binary information sequences, which are modulated to complex symbols, are transmitted by users over a flat fading channel. Using a discrete time model, $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$ is the received symbol vector written as:

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{1}$$

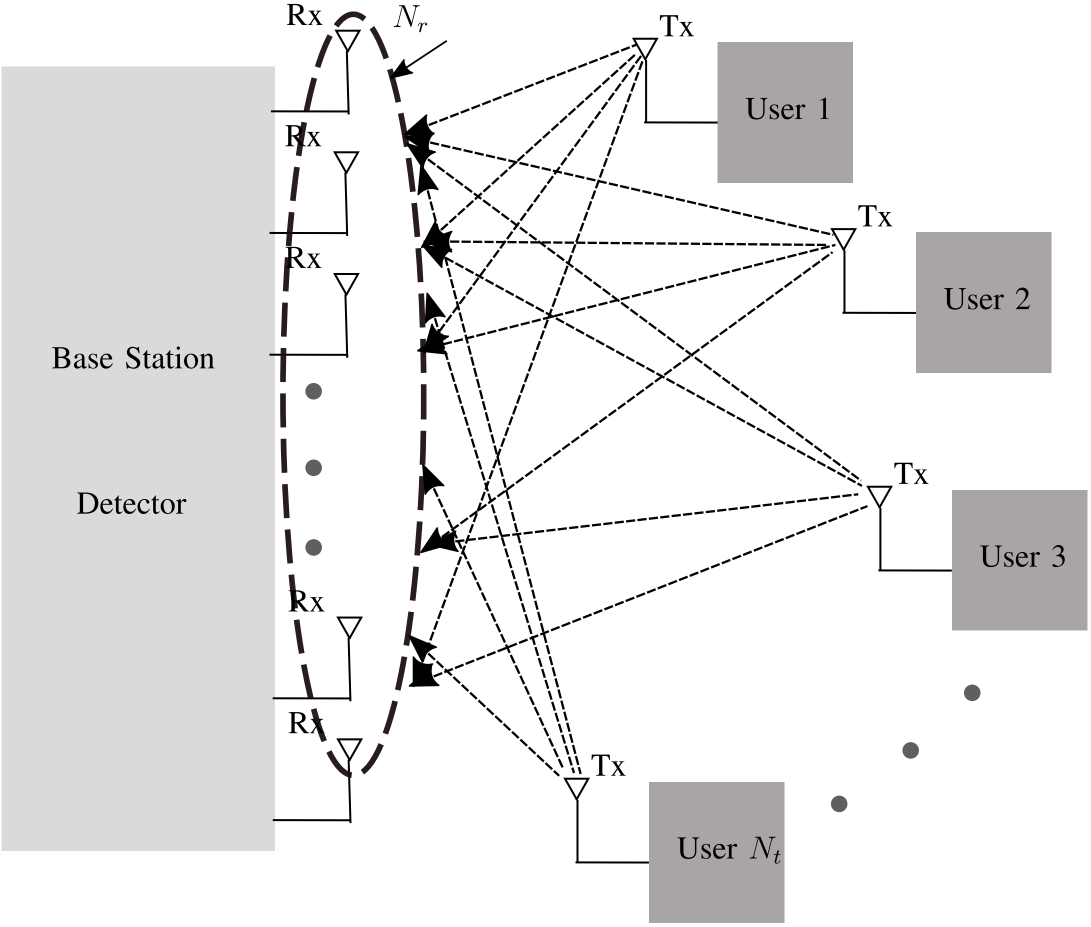

Fig. 1. Large MIMO uplink system

where $\mathbf{s} \in \mathbb{C}^{N_t}$ is the transmitted symbol vector, with components that are mutually independent and taken from a finite signal constellation alphabet $\mathbb{O}$ (e.g. 4-QAM, 16-QAM, 64-QAM) of size M. The possible 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, and $\mathbb{E}[\cdot]$ denotes the expectation operation. Furthermore $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ denotes the Rayleigh fading channel propagation matrix with independent identically distributed (i.i.d) circularly symmetric complex Gaussian zero mean components with 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).

Assume the receiver has perfect channel state information (CSI), meaning that $\mathbf{H}$ is known, as well as the SNR. The task of the MIMO decoder is to recover $\mathbf{s}$ based on $\mathbf{y}$ and $\mathbf{H}$.

III. Brief Introduction to ϵ-Support Vector Regression

A. Regression Model

Suppose we are given training data set $((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_l, y_l))$, l denotes the number of training samples, $\mathbf{x} \in \mathbb{R}^v$ denotes input data vector, v is the number of features in $\mathbf{x}$. y denotes output. The regression model (either linear or non-linear regression) is given by

$$y_i = \mathbf{w}^T \Phi(\mathbf{x}_i) + b \quad i \in 1 \cdots l \tag{2}$$

where $\mathbf{w}$ denotes regression coefficient vector, $\Phi(x)$ denotes the mapping of $\mathbf{x}$ to higher dimensional feature space. 2.

Here we give the primal optimization problem directly

$$\frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{l} C_i(R(\xi_i) + R(\hat{\xi}_i))$$

$$s.t. \begin{cases} y_i - \mathbf{w}^T \Phi(\mathbf{x}_i) - b & \leq \epsilon + \xi_i \\ \mathbf{w}^T \Phi(\mathbf{x}_i) + b - y_i & \leq \epsilon + \hat{\xi}_i \\ \epsilon, \xi, \hat{\xi} & \geq 0 \end{cases} \tag{3}$$

In 3, $\frac{1}{2}||\mathbf{w}||^2$ is the regularization term in order to ensure the flatness of regression model. ϵ denotes the precision, if the error between estimation and real output is less than ϵ, As shown in

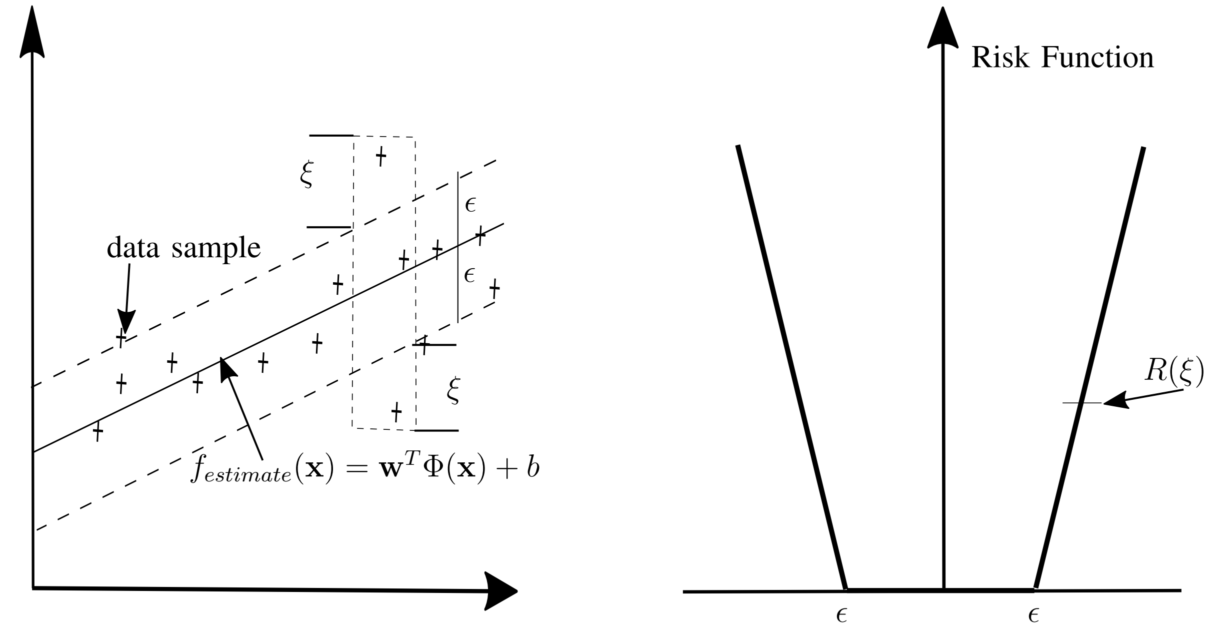

Fig. 2. ϵ-Support Vector Regression and Risk Functional

Fig 2, only those data points outside the shadow part, which is called ϵ tube, contribute to cost function. ξ and $\hat{\xi}$ denote slack variables that cope with noise of input data set, $R(x)$ denotes risk function, the simplest risk function is $R(x) = x$, risk function is determined by the statistical distribution of noise [**?**], for example if the noise subject to Gaussian distribution, the optimal cost function is $R(x) = \frac{1}{2}x^2$. $C\sum_{i=1}^{l}(\xi_i + \hat{\xi}_i)$ denotes the penalty of noise, $C \in \mathbb{R}$ and $C \geq 0$ controls the trade off between regularization term and noise penalty term.

B. Risk Functional

From the rationale of regularized risk function, let $f_{true}(\mathbf{x})$ denotes true regression function and $f_{estimate}(\mathbf{x})$, $c(\mathbf{x}, y, f_{estimate}(\mathbf{x}))$ denotes the risk function, the regression model can be written as $y = f_{true}(\mathbf{x}) + \xi$, ξ denotes additive noise. Assume the data samples are i.i.d. Based on Maximum Likelihood (ML) principle we want to

$$\begin{aligned} maximize \quad \prod_{i=1}^{l} P(y_i|f_{estimation}(\mathbf{x}_i)) \qquad &= maximize \quad \textstyle\prod_{i=1}^{l} P(\xi_i) \\ &= maximize \quad \textstyle\prod_{i=1}^{l} P(y_i - f(\mathbf{x}_i)), \end{aligned} \tag{4}$$

Take the logarithm of (79), we have

$$maximize \quad \sum_{i=1}^{l} log(P(y_i - f_{estimation}(\mathbf{x}_i))), \tag{5}$$

Therefore the ith risk function of $(\mathbf{x_i}, y_i)$ can be written as

$$c(\mathbf{x}_i, y_i, f_{estimation}(\mathbf{x}_i)) = -log(P(y_i - f_{estimation}(\mathbf{x}_i))). \tag{6}$$

Thus the equivalent formula of (12) can be written as

$$minimize \quad \sum_{i=1}^{l} c(\mathbf{x}_i, y_i, f_{estimation}(\mathbf{x}_i)), \tag{7}$$

In ϵ-SVR, Vapnik's ϵ-insensitive function, as shown in (8), is applied to (6).

$$|x|_\epsilon = \begin{cases} 0 & if \quad |x| < \epsilon \\ |x| - \epsilon & otherwise \end{cases} \tag{8}$$

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

$$\tilde{c}(\mathbf{x}, y, f_{estimation}(\mathbf{x})) = \frac{1}{l}\sum_{i=1}^{l} m_i(-log(P(|y_i - f_{estimation}(\mathbf{x}_i)|_\epsilon))), \tag{9}$$

where $m_i \in \mathbb{R}$, $m_i > 0$ denotes the weight parameter, if $y_i > f_{estiamtion}(\mathbf{x})$, $m_i = m_{positive}$, else $m_i = m_{negative}$, Therefore the regularized risk function can written as

$$minimize \quad \lambda||\mathbf{w}||^2 + \tilde{c}(\mathbf{x}, y, f_{estimation}(\mathbf{x})), \tag{10}$$

where λ denotes the weight of regularization term, divide (10) by $\frac{1}{2\lambda}$, we have the optimization problem

$$minimize \quad \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{l} C_i(-log(P(|y_i - f_{estimation}(\mathbf{x}_i)|_\epsilon))), \tag{11}$$

where $C_i = \frac{m_i}{2\lambda l}$, based on (11), by introducing slack variables, we can easily derive the equivalent optimization problem as same as (3):

$$\begin{aligned} minimize \quad & f(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{l} C_i(R(\xi_i) + R(\hat{\xi}_i)) \\ s.t. & \begin{cases} y_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b & \leq \epsilon + \xi_i \\ \mathbf{w}^T\Phi(\mathbf{x}_i) + b - y_i & \leq \epsilon + \hat{\xi}_i \\ \epsilon, \xi, \hat{\xi} & \geq 0 \end{cases} \end{aligned} \tag{12}$$

where $R(x) = -log(P(x))$, by this way, the discontinuity of ϵ-insensitive function is conquered, we arrive to at a convex minimization problem [**?**].

C. Lagrange Duality

According to Lagrange Theorem, the optimization problem (12) can be transferred to dual form by combining original objective function with linear combination of equality and inequality constraints, the combination coefficient is called Lagrange multiplier. Thus we have Lagrange function.

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

where η,$\hat{\eta}$, α, $\hat{\alpha}$ are Lagrange multipliers.

The sufficient and necessary condition that $f(\mathbf{w}^*)$ is the global minimum of (12) is called Karush-Kuhn-Tucker (KKT) conditions, which can be written as

$$\frac{\partial\Theta}{\partial\mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l}(\alpha_i - \hat{\alpha}_i)\Phi(\mathbf{x}_i) = 0 \tag{14}$$

$$\frac{\partial\Theta}{\partial\xi} = C_iR^{'}(\xi_i) - \eta_i - \alpha_i = 0 \tag{15}$$

$$\frac{\partial\Theta}{\partial\hat{\xi}} = C_iR^{'}(\hat{\xi}_i) - \hat{\eta}_i - \hat{\alpha}_i = 0 \tag{16}$$

$$\frac{\partial\Theta}{\partial b} = \sum_{i=1}^{l}(\alpha_i - \hat{\alpha}_i) = 0 \tag{17}$$

$$\begin{cases} \alpha, \hat{\alpha} \geq 0 \\ \alpha(y - \mathbf{w}^T\Phi(\mathbf{x}) - b - \epsilon - \xi) = 0 \\ \hat{\alpha}(\mathbf{w}^T\Phi(\mathbf{x}) + b - y - \epsilon - \hat{\xi}) = 0 \end{cases} \tag{18}$$

The conditions in (18) is called KKT complimentary condition. Then substitute (15)-(18) to (13) and for sake of brevity, we make C_i uniform to all data samples, we have the dual form of objective function

$$\theta(\alpha, \hat{\alpha}) = \frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}(\alpha_i - \hat{\alpha_i})(\alpha_j - \hat{\alpha_j})\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i) + C\sum_{i=1}^{l}[(R(\xi_i) - \xi_i R^{'}(\xi_i))$$

$$+(R(\hat{\xi_i}) - \hat{\xi_i} R^{'}(\hat{\xi_i}))] + \sum_{i=1}^{l}[(\alpha_i - \hat{\alpha_i})y_i - (\alpha_i + \hat{\alpha_i})\epsilon]$$

$$-\sum_{i=1}^{l}\sum_{i=1}^{l}(\alpha_i - \hat{\alpha_i})(\alpha_j - \hat{\alpha_j})\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i),$$

$$s.t. \begin{cases} \sum_{i=1}^{l}(\alpha_i - \hat{\alpha_i}) = 0 \\ 0 < \alpha < C\tilde{R}^{'}(\alpha) \\ 0 < \hat{\alpha} < C\tilde{R}^{'}(\hat{\alpha}) \end{cases} \tag{19}$$

Obviously $\theta(\alpha, \hat{\alpha}) \leq \Theta(\mathbf{w}, \alpha, \hat{\alpha}, \eta, \hat{\eta})$, notice now (15)-(18)are satisfied, therefore, KKT complimentary condition (18) is the only requirement to find global optimal point of original optimization problem. Because

$$\Theta(\mathbf{w}, \alpha, \hat{\alpha}, \eta, \hat{\eta}) = f(\mathbf{w}) + \sum_{i}(\alpha_i g_i(\mathbf{w}) + \hat{\alpha}_i \hat{g}_i(\mathbf{w})) + \sum_{i}(\eta_i l_i(\mathbf{w}) + \hat{\eta}_i \hat{l}_i(\mathbf{w})), \tag{20}$$

where $g_i(\mathbf{w})$, $g_i(\hat{\mathbf{w}})$, $l_i(\mathbf{w})$, $\hat{l}_i(\mathbf{w})$ denote inequality constraints.

$$g_i(\mathbf{w}) = \mathbf{y}_i - \mathbf{w}^T\Phi(\mathbf{x}_i) - b - \epsilon - \xi_i \leq 0 \tag{21}$$

$$\hat{g}_i(\mathbf{w}) = \mathbf{w}^T\Phi(\mathbf{x}_i) + b - \mathbf{y}_i - \epsilon - \hat{\xi}_i \leq 0 \tag{22}$$

$$l_i(\mathbf{w}) = -\xi_i \leq 0 \tag{23}$$

$$\hat{l}_i(\mathbf{w}) = -\hat{\xi}_i \leq 0 \tag{24}$$

Hence $\Theta \leq f(\mathbf{w})$, we have $\theta \leq \Theta \leq f(\mathbf{w})$, therefore the upper bound of θ is determined by the original objective function $f(\mathbf{w})$. when $\theta = f(\mathbf{w})$, according to (20), the linear combination term of inequality constraints equal to zero, that is, KKT complimentary conditions in (18) satisfied. hence the global optimal point is found for both θ and $f(\mathbf{w})$ if and only if the equality holds. Therefore duality gap is defined as $G = f(\mathbf{w}) - \theta$, which can be used as an evaluation for the closeness of one solution to the global optimal.

In conclusion, the dual objective function can be written as

$$\begin{aligned}
&maximize \quad \theta = -\frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i) + \sum_{i=i}^{l}[(\alpha_i - \hat{\alpha}_i)y_i - (\alpha_i + \hat{\alpha}_i)\epsilon] \\
&+C\sum_{i=1}^{l}[\tilde{R}(\xi_i) + \tilde{R}(\hat{\xi}_i)] \\
&= -\frac{1}{2}(\mathbf{a} - \hat{\mathbf{a}})^T\mathbf{K}(\mathbf{a} - \hat{\mathbf{a}}) + (\mathbf{y} - \epsilon)^T\mathbf{a} + (-\mathbf{y} - \epsilon)^T\hat{\mathbf{a}} + \mathbf{e}^T C(\tilde{R}(\xi) + \tilde{R}(\hat{\xi})),
\end{aligned} \tag{25}$$

where $\mathbf{a} = [\alpha_1, \alpha_2, \ldots, \alpha_l]^T$, $\hat{\mathbf{a}} = [\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_l]^T$, $\mathbf{y} = [y_1, y_2, \ldots y_l]^T$, $\mathbf{e} = [1, 1, \ldots, 1]^T \in \mathbb{R}^l$, $\mathbf{e}_i$ denotes the vector that only ith component is 1 while the rest are all 0, $\tilde{R}(\xi) = R(\xi) - \xi R'(\xi) \in$

$\mathbb{R}^l$, $\mathbf{K}_{ij} = \Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i)$ denotes data kernel matrix. We define the following $2l$ vectors $\mathbf{a}^{(*)}$= $\left[\begin{smallmatrix}\mathbf{a}\\ \hat{\mathbf{a}}\end{smallmatrix}\right]$, $\mathbf{v} \in \mathbb{R}^{2l}$,

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

(25) can also be reformulate as

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

IV. Dual Channel Real Kernel Complex Support Vector Regression for Large MIMO system

Based on discrete time model of large MIMO uplink system in (1), in our regression model, the training data sample at detector is $(\mathbf{h}_1, y_1)(\mathbf{h}_2, y_2), \dots, (\mathbf{h}_{N_r}, y_{N_r})$, where $\mathbf{h}_i$ denotes ith row of channel propagation matrix $\mathbf{H}$, this yields a regression task without offset b:

$$y_i = f_{true}(\mathbf{h}_i) + n, \tag{28}$$

$$f_{true}(\mathbf{h}_i) = \mathbf{h}_i\mathbf{s}, \tag{29}$$

$$\tag{30}$$

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

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

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

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

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

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

$$\mathbf{a}^T\mathbf{b}\in<\mathbf{a},\mathbf{b}>_H, \tag{31}$$

From Property 1 and Property 3, it is obvious

$$<\mathbf{g},\mathbf{h}>_{\mathbb{H}}=<\mathbf{g}^r,\mathbf{h}^r>_H+<\mathbf{g}^i,\mathbf{h}^i>_H+i(<\mathbf{g}^i,\mathbf{h}^r>_H-<\mathbf{g}^r,\mathbf{h}^i>_H) \tag{32}$$

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

$$\begin{aligned} < \mathbf{h}, \mathbf{s}^* >_{\mathbb{H}} &= < \mathbf{h}^r, (\mathbf{s}^*)^r >_H + < \mathbf{h}^i, (\mathbf{s}^*)^i >_H + i(< \mathbf{h}^i, (\mathbf{s}^*)^r >_H - < \mathbf{h}^r, (\mathbf{s}^*)^i >_H) \\ &= < \mathbf{h}^r, \mathbf{s}^r >_H - < \mathbf{h}^i, \mathbf{s}^i >_H + i(< \mathbf{h}^i, \mathbf{s}^r >_H + < \mathbf{h}^r, \mathbf{s}^i >_H), \qquad (33) \end{aligned}$$

$$\mathbf{h}\mathbf{s} = \mathbf{h}^r\mathbf{s}^r - \mathbf{h}^i\mathbf{s}^i + i(\mathbf{h}^i\mathbf{s}^r + \mathbf{h}^r\mathbf{s}^i), \qquad (34)$$

Because of (31), (33) and (34), $\mathbf{h}_i\mathbf{s} \in < \mathbf{h}_i, \mathbf{s}^* >_{\mathbb{H}}$ □

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

$$\begin{aligned} &minimize \quad \frac{1}{2}||w||_{\mathbb{H}}^2 + C\sum_{k=1}^{N_r}[(R(\xi_k^r) + R(\hat{\xi}_k^r) + R(\xi_k^i) + R(\hat{\xi}_k^i))] \\ &s.t. \begin{cases} Re(y_k - < \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) \le \epsilon + \xi_k^r \\ Re(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}} - y_k) \le \epsilon + \hat{\xi}_k^r \\ Im(y_k - < \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) \le \epsilon + \xi_k^i \\ Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}} - y_k) \le \epsilon + \hat{\xi}_k^i \\ \xi^r, \hat{\xi}^r, \xi^i, \hat{\xi}^i \ge 0 \end{cases} \qquad (35) \end{aligned}$$

where $Re()$ and $Im()$ denote real part and imaginary part of a complex variable, restrictions are set to real and imaginary part of regression function separately. Let $\mathbf{K} = \mathbf{H}\mathbf{H}^H$ denotes the kernel function, $\mathbf{K} = \mathbf{K}^r + i\mathbf{K}^i$, $\mathbf{K}^r$ and $\mathbf{K}^i$ denote matrix of corresponding real part and imaginary part. Similar to the Lagrange duality rational in (**??**), Lagrange duality is formulated for (35)

$$\max_{(\alpha,\hat{\alpha},\beta,\hat{\beta},\eta,\hat{\eta},\tau,\hat{\tau})} \min_{(\mathbf{w},\xi^r,\hat{\xi}^r,\hat{\xi}^i,\hat{\xi}^i)} \theta = \frac{1}{2}||w||^2_{\mathbb{H}} + C\sum_{k=1}^{N_r}[(R(\xi_k^r) + R(\hat{\xi}_k^r) + R(\xi_k^i) + R(\hat{\xi}_k^i))] - \sum_{k=1}^{N_r}(\eta_i\xi_k^r + \hat{\eta}_k\hat{\xi}_k^r$$

$$+\tau_k\xi_k^i + \hat{\tau}_k\hat{\xi}_k^i) - \sum_{k=1}^{N_r}\alpha_k(\epsilon + \xi_k^r - Re(y_k) + Re(<\mathbf{h}_k, \mathbf{w}>_{\mathbb{H}})) - \sum_{k=1}^{N_r}\hat{\alpha}_k(\epsilon + \hat{\xi}_k^r + Re(y_k) - Re(<\mathbf{h}_k, \mathbf{w}>_{\mathbb{H}}))$$

$$-\sum_{k=1}^{N_r}\beta_k(\epsilon + \xi_k^i - Im(y_k) + Im(<\mathbf{h}_k, \mathbf{w}>_{\mathbb{H}})) - \sum_{k=1}^{N_r}\hat{\beta}_k(\epsilon + \hat{\xi}_k^i + Im(y_k) - Im(<\mathbf{h}_k, \mathbf{w}>_{\mathbb{H}}))$$

$$s.t. \begin{cases} \eta, \hat{\eta}, \tau, \hat{\tau}\alpha, \hat{\alpha}, \beta, \hat{\beta} \geq 0 \\ \xi^r, \hat{\xi}^r, \xi^i, \hat{\xi}^i \geq 0 \end{cases} \tag{36}$$

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

$$\begin{cases} \frac{\partial\Theta}{\partial\mathbf{w}^*} = \frac{1}{2}\mathbf{w} - \frac{1}{2}\sum_{k=1}^{N_r}\alpha_k\mathbf{h}_k + \frac{1}{2}\sum_{k=1}^{N_r}\hat{\alpha}_k\mathbf{h}_k + \frac{i}{2}(\sum_{k=1}^{N_r}\beta_k\mathbf{h}_k - \sum_{k=1}^{N_r}\hat{\beta}_k\mathbf{h}_k) = 0 \\ \Rightarrow \mathbf{w} = \sum_{k=1}^{N_r}(\alpha_k - \hat{\alpha}_k)\mathbf{h}_k - i\sum_{k=1}^{N_r}(\beta_k - \hat{\beta}_k)\mathbf{h}_k \\ \frac{\partial\Theta}{\partial\xi_k^r} = CR^{'}(\xi_k^r) - \eta_k - \alpha_k = 0 \Rightarrow \eta_k = CR^{'}(\xi_k^r) - \alpha_k \\ \frac{\partial\Theta}{\partial\hat{\xi}_k^r} = CR^{'}(\hat{\xi}_k^r) - \hat{\eta}_k - \hat{\alpha}_k = 0 \Rightarrow \hat{\eta}_k = CR^{'}(\hat{\xi}_k^r) - \hat{\alpha}_k \\ \frac{\partial\Theta}{\partial\xi_k^i} = CR^{'}(\xi_k^i) - \tau_k - \beta_k = 0 \Rightarrow \tau_k = CR^{'}(\xi_k^i) - \beta_k \\ \frac{\partial\Theta}{\partial\hat{\xi}_k^i} = CR^{'}(\hat{\xi}_k^i) - \hat{\eta}_k - \hat{\beta}_k = 0 \Rightarrow \hat{\eta}_k = CR^{'}(\hat{\xi}_k^i) - \hat{\beta}_k \end{cases} \tag{37}$$

Based on (37), we have

$$
\begin{aligned}
&< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}= \sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} +i \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} \\
&= \sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) \mathbf{h}_i \mathbf{h}_j^H + i \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) \mathbf{h}_i \mathbf{h}_j^H \\
&= \sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) \mathbf{K}_{ij}^r - \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) \mathbf{K}_{ij}^i + i (\sum_{j=1}^{N_r} (\alpha_j - \hat{\alpha}_j) \mathbf{K}_{ij}^i + \sum_{j=1}^{N_r} (\beta_j - \hat{\beta}_j) \mathbf{K}_{ij}^r), \quad (38)
\end{aligned}
$$

$$
\begin{aligned}
||\mathbf{w}||_{\mathbb{H}}^2 &= \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\alpha_i - \hat{\alpha}_i) \mathbf{h}_i \mathbf{h}_j^H + \sum_{i,j=1}^{N_r} (\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_i) \mathbf{h}_i \mathbf{h}_j^H \\
&+ i (\sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j) \mathbf{h}_i \mathbf{h}_j^H - \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j) \mathbf{h}_j \mathbf{h}_i^H) \quad (39)
\end{aligned}
$$

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

$$
\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i = - \sum_{i,j}^{l} r_i r_j \mathbf{K}_{ji}^i = - \sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i, \quad (40)
$$

Therefore

$$
\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i = 0, \quad (41)
$$

Based on (41), (39) can be changed to

$$
||\mathbf{w}||_{\mathbb{H}}^2 = \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\alpha_i - \hat{\alpha}_i) \mathbf{K}_{ij}^r + \sum_{i,j=1}^{N_r} (\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_i) \mathbf{K}_{ij}^r - 2 \sum_{i,j=1}^{N_r} (\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j) \mathbf{K}_{ij}^i. \quad (42)
$$

Apply (37), (38), (41) and (42) to (36), the final form of Lagrange duality can be obtained

$$
\begin{aligned}
&maximize \quad \theta = -\frac{1}{2}[\sum_{i,j}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{ij}^r + \sum_{i,j}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)\mathbf{K}_{ij}^r] \\
&-\sum_{i}^{N_r}(\alpha_i + \hat{\alpha}_i + \beta + \hat{\beta}_i)\epsilon + [\sum_{i=1}^{N_r}(\alpha_i - \hat{\alpha}_i)Re(y_i) + \sum_{i=1}^{N_r}(\beta_i - \hat{\beta}_i)Im(y_i)] \\
&+C\sum_{i}^{N_r}(\tilde{R}(\xi_i^r) + \tilde{R}(\hat{\xi}_i^r) + \tilde{R}(\xi_i^i) + \tilde{R}(\hat{\xi}_i^i)) \\
&\begin{cases}
0 \leq \alpha(\hat{\alpha}) \leq C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r)) \\
0 \leq \beta(\hat{\beta}) \leq C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i)) \\
\xi^r(\hat{\xi}^r) \geq 0 \\
\xi^i(\hat{\xi}^i) \geq 0
\end{cases}
\end{aligned}
\tag{43}
$$

which can be divided into 2 independent regression task,

$$
\begin{aligned}
&maximize \quad \Theta^r = -\frac{1}{2}\sum_{i,j}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{ij}^r - \sum_{i=1}^{N_r}(\alpha_i + \hat{\alpha}_i)\epsilon + \sum_{i=1}^{N_r}(\alpha_i - \hat{\alpha}_i)Re(y_i) + C\sum_{i=1}^{N_r}(\tilde{R}(\xi_i^r) \\
&+\tilde{R}(\hat{\xi}_i^r)) \\
&\begin{cases}
0 \leq \alpha(\hat{\alpha}) \leq C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r)) \\
\xi^r(\hat{\xi}^r) \geq 0
\end{cases}
\end{aligned}
\tag{44}
$$

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

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

$$\begin{cases} 0 \leq \beta(\hat{\beta}) \leq C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i)) \\ \xi^i(\hat{\xi}^i) \geq 0 \end{cases} \tag{45}$$

The alternate form can be written as

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

$$\begin{cases} 0 \leq \alpha(\hat{\alpha}) \leq C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r)) \\ \xi^r(\hat{\xi}^r) \geq 0 \end{cases} \tag{46}$$

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

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

where $(\alpha - \hat{\alpha})$, $(\beta - \hat{\beta})$,$Re(\mathbf{y})$, $Im(\mathbf{y})$ denote vectors, $\mathbf{e} = [1, 1, \ldots, 1]^T \in \mathbb{R}^{N_r}$, $\mathbf{K}^r$ denotes the matrix consist of real part of kernel components. Observe that solving (46) and (47) are equivalent to solving two independent real Support vector regression task (dual channel), only the real part of kernel matrix is required for each channel. In section VII, we will further show that from the statistic analyst of channel orthogonality (which is also named channel hardening phenomenon), the imaginary part of kernel matrix can also be omitted in stopping criteria.

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

V. Work Set Selection and Solver

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

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

Because (44) and (45) are symmetric, in this section we discuss real part only. By dividing the variables into work set S and fixed set N, we have $(\alpha, \hat{\alpha}) = ((\alpha_N, \hat{\alpha}_N)\mathbf{e}_N + (\alpha_S + \hat{\alpha}_S)\mathbf{e}_S)$, where $\mathbf{e}_K$ denotes the modified vector of $\mathbf{e}$ with the components in set K zeroed, for sake of brevity we replace $\alpha_K \mathbf{e}_K$ by α_K. Thus (46) can be changed to:

$$
\begin{aligned}
&maximize \quad \Theta^r = -\frac{1}{2}[(\alpha-\hat{\alpha})_S^T \mathbf{K}_{SS}^r (\alpha-\hat{\alpha})_S + 2(\alpha-\hat{\alpha})_N^T \mathbf{K}_{NS}^r (\alpha-\hat{\alpha})_S] + Re(\mathbf{y})_S^T (\alpha-\hat{\alpha})_S - \\
&\epsilon(\mathbf{e}^T(\alpha+\hat{\alpha})_S) - \frac{1}{2}(\alpha-\hat{\alpha})_N^T \mathbf{K}_{NN}^r (\alpha-\hat{\alpha})_N + Re(\mathbf{y})_N^T (\alpha-\hat{\alpha})_N - \epsilon(\mathbf{e}^T(\alpha+\hat{\alpha})_N) \\
&+C(\mathbf{e}^T(\tilde{R}(\xi^r) + \tilde{R}(\hat{\xi}^r))),
\end{aligned} \tag{48}
$$

Where $\mathbf{K}^r = \begin{bmatrix} \mathbf{K}^r_{SS} & \mathbf{K}^r_{SN} \\ \mathbf{K}^r_{NS} & \mathbf{K}^r_{NN} \end{bmatrix}$ is a permutation of $\mathbf{K}^r$, $\mathbf{K}^r_{SN} = \mathbf{K}^r_{NS}$ and $\alpha_S \in \mathbb{R}^{N_r}$ denotes the vector with the components that do not belong to S zeroed. In each iteration, in (48), α_N is fixed and only the sub problem that correlated to α_S is solved i.e

$$maximize \quad \Theta^r_S = -\frac{1}{2}[(\alpha-\hat{\alpha})^T_S \mathbf{K}^r_{SS}(\alpha-\hat{\alpha})_S]+[Re(\mathbf{y})^T_S-(\alpha-\hat{\alpha})^T_N \mathbf{K}^r_{NS}](\alpha-\hat{\alpha})_S-\epsilon[\mathbf{e}^T_S(\alpha+\hat{\alpha})_S], \tag{49}$$

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

Another class of approaches is to choose the dual variables whose update can provide the maximum improvements to objective function [], [], [](Training without offset). That is

$$maximize \quad \bigtriangledown \Theta_S = \Theta_S((\alpha_S + \delta_S \mathbf{e}_S), (\hat{\alpha}_s + \hat{\delta}_S \mathbf{e}_S)) - \Theta_S(\alpha_S, \hat{\alpha}_S), \tag{50}$$

where $\delta_S = \alpha_S^{new} - \alpha_S$, the gain in (50) can be written as

$$\begin{aligned}&\bigtriangledown\Theta_S^r = -\frac{1}{2}[(\delta-\hat{\delta})_S^T\mathbf{K}_{SS}^r(\delta-\hat{\delta})_S + 2(\alpha-\hat{\alpha})_S^T\mathbf{K}_{SS}^r(\delta-\hat{\delta})_S] + [Re(\mathbf{y})_S^T - (\alpha-\hat{\alpha})_N^T\mathbf{K}_{NS}^r](\delta-\hat{\delta})_S \\ &-\epsilon\mathbf{e}_S^T(\delta+\hat{\delta})_S = -\frac{1}{2}(\delta-\hat{\delta})_S^T\mathbf{K}_{SS}^r(\delta-\hat{\delta})_S + [Re(\mathbf{y})_S^T - (\alpha-\hat{\alpha})^T\mathbf{K}_S^r](\delta-\hat{\delta})_S - \epsilon\mathbf{e}_S^T(\delta+\hat{\delta})_S \quad (51)\end{aligned}$$

In (51), we use

$$(\alpha-\hat{\alpha})_S^T\mathbf{K}_{SS}^r + (\alpha-\hat{\alpha})_N^T\mathbf{K}_{NS}^r = [(\alpha-\hat{\alpha})_S^T, (\alpha-\hat{\alpha})_N^T]\begin{bmatrix}\mathbf{K}_{SS}^r \\ \mathbf{K}_{NS}^r\end{bmatrix} = (\alpha-\hat{\alpha})^T\mathbf{K}_S^r, \qquad (52)$$

where $\mathbf{K}_S^r \in \mathbb{R}^{N_r \times S}$ denotes the matrix constructed by all the columns that belong to work set S. Then we define intermediate variable vector $\Phi \in \mathbb{C}^{N_r}$, $\Phi^r = Re(\mathbf{y}) - \mathbf{K}^r(\alpha - \hat{\alpha})$ and $\Phi^i = Im(\mathbf{y}) - \mathbf{K}^r(\beta - \hat{\beta})$ Thus (51) can be rewritten as

$$\bigtriangledown\Theta_S^r = -\frac{1}{2}(\delta-\hat{\delta})_S^T\mathbf{K}_{SS}^r(\delta-\hat{\delta})_S + (\Phi_S^r)^T(\delta-\hat{\delta})_S - \epsilon\mathbf{e}_S^T(\delta+\hat{\delta})_S \qquad (53)$$

The offset term is omitted in Large MIMO system, therefore different from SMO type algorithms, there is no linear equation constraint as shown in (19), it is possible to update only one variable pair in each iteration. However, recent work shows more efficient work set selection strategy based on maximum gain selection approaches, that choose two pair of dual variables can reduce computational cost while maintaining the comparable performance with that with offset [**?**]. Here we propose sequential 1-D work set selection strategy, which can approximate the performance of optimal 2-D work set selection, while only $O(n)$ searching times required for the former one instead of $O(n^2)$ searching times.

A. Single Direction Solver

We will first introduce 1-D work set selection strategy in which the dual variable pair that maximizes the gain of objective function is chosen. Recall the decomposition method in (50), let α_1 denotes the dual variable that is chosen to be updated. The sub optimization problem is formulated as

$$maximize \quad \Theta_1^r = -\frac{1}{2}((\alpha_1 - \hat{\alpha}_1)^{new})^2 \mathbf{K}_{11}^r - (\alpha_1 - \hat{\alpha}_1)^{new} \sum_{j=2}^{N_r} \mathbf{K}_{1j}^r (\alpha_j - \hat{\alpha}_j) + Re(y_1)(\alpha_1 - \hat{\alpha}_1)^{new}$$
$$-\epsilon(\alpha_1 + \hat{\alpha}_1)^{new}, \tag{54}$$

take the partial derivative of Θ_1^r respect to α_1 and $\hat{\alpha}_1$, we have

$$\begin{aligned} &\frac{\partial \Theta_1^r}{\partial \alpha_1} = -(\alpha_1 - \hat{\alpha}_1)^{new} \mathbf{K}_{11}^r - \sum_{j=2}^{N_r} (\alpha_j - \hat{\alpha}_j)^{new} \mathbf{K}_{1j}^r + Re(y_1) - \epsilon = 0 \\ &\Rightarrow \alpha_1^{new} = \alpha_1 + \frac{\Phi_1^r - \epsilon}{\mathbf{K}_{11}^r}, \end{aligned} \tag{55}$$

$$\begin{aligned} &\frac{\partial \Theta_1^r}{\partial \hat{\alpha}_1} = (\alpha_1 - \hat{\alpha}_1)^{new} \mathbf{K}_{11}^r + \sum_{j=2}^{N_r} (\alpha_j - \hat{\alpha}_j)^{new} \mathbf{K}_{1j}^r - Re(y_1) - \epsilon = 0 \\ &\Rightarrow \hat{\alpha}_1^{new} = \hat{\alpha}_1 - \frac{\Phi_1^r + \epsilon}{\mathbf{K}_{11}^r} \end{aligned} \tag{56}$$

where we define $\Phi_i^r = Re(y_i) - \sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{ij}^r$, similarly, as to dual variable β and $\hat{\beta}$, we define $\Phi_i^i = Im(y_i) - \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\mathbf{K}_{ij}^r$. Recall complementary KKT condition

$$\begin{cases} (C\tilde{R}(\xi^r) - \alpha)\xi^r = 0 \\ (C\tilde{R}(\hat{\xi}^r) - \hat{\alpha})\hat{\xi}^r = 0 \\ \alpha(\epsilon + \xi^r - Re(y) + < \mathbf{h}, \mathbf{w} >_{\mathbb{H}}) = 0 \\ \hat{\alpha}(\epsilon + \hat{\xi}^r + Re(y) - < \mathbf{h}, \mathbf{w} >_{\mathbb{H}}) = 0 \end{cases} \tag{57}$$

it can be easily observed that $\alpha\hat{\alpha} = 0$, because $0 \leq \alpha(\hat{\alpha}) \leq C\tilde{R}(\xi^r)(C\tilde{R}(\hat{\xi}^r))$, ξ^r and $\hat{\xi}^r$ satisfy $\xi^r\hat{\xi}^r$. The update of α_1 or $\hat{\alpha}_1$ is completed by clipping

$$\alpha^{new \quad clipped} = [\alpha^{new}]_0^{C\tilde{R}(\xi^r)} \tag{58}$$

$$\hat{\alpha}^{new \quad clipped} = [\hat{\alpha}^{new}]_0^{C\tilde{R}(\hat{\xi}^r)} \tag{59}$$

where $[]_a^b$ denotes clipping function

$$[x]_a^b = \begin{cases} a \quad if \quad x \leq a \\ x \quad if \quad a < x < b \\ b \quad if \quad x \geq b \end{cases} \tag{60}$$

Based on (53), The gain of objective function respect to ith dual variable pair is

$$\begin{aligned}\bigtriangledown\Theta_i^r &= \Theta^r((\alpha_i+\delta_i\mathbf{e}_i),(\hat{\alpha}_i+\hat{\delta}_i\mathbf{e}_i)) - \Theta^r(\alpha,\hat{\alpha}) \\ &= -\frac{1}{2}(\delta_i-\hat{\delta}_i)^2\mathbf{K}_{ii}^r + \Phi_1^r(\delta_i-\hat{\delta}_i) - \epsilon(\delta_i+\hat{\delta}_i) \\ &= (\delta_i-\hat{\delta}_i)[-\frac{1}{2}(\delta_i-\hat{\delta}_i)\mathbf{K}_{ii}^r + \Phi_i^r - \epsilon\frac{\delta_i+\hat{\delta}_i}{\delta_i-\hat{\delta}_i}],\end{aligned} \tag{61}$$

where $\delta_1 = \alpha_1^{new \quad clipped} - \alpha_1$, $\hat{\delta}_1 = \hat{\alpha}_1^{new \quad clipped} - \hat{\alpha}_1$, in 1-D searching procedure, the dual variable pair which has the maximum gain of objective function is chosen as 1 in (54), that is

$$1 = arg_{(i=1,\ldots,N_r)} \max \bigtriangledown\Theta_i^r, \tag{62}$$

B. Double Direction Solver

Although omission of offset in the CSVR-MIMO detector makes 1-D solver possible, however recent work in machine learning field shows training SVM without offset by 2-D solver with special work set selection strategies has more rapid training speed while the comparable performance is retained [**?**]. The 2-D solver uses the same principle as 1-D solver, the work set size is 2, recall (49), let $(\alpha_1,\hat{\alpha}_1)$, $(\alpha_2,\hat{\alpha}_2)$ denote the two dual variable pairs that are chosen for optimization, that is $(\alpha_s,\hat{\alpha}_s) = ((\alpha_1\mathbf{e}_1+\alpha_2\mathbf{e}_2),(\hat{\alpha}_1\mathbf{e}_1+\hat{\alpha}_2\mathbf{e}_2))$. Thus we have, based on (49), the sub objective function can be written as

$$\begin{aligned}&maximize \quad \Theta_{1,2}^r = -\frac{1}{2}[(\alpha_1-\hat{\alpha}_1)^2\mathbf{K}_{11}^r + (\alpha_2-\hat{\alpha}_2)^2\mathbf{K}_{22}^r + 2(\alpha_1-\hat{\alpha}_1)(\alpha_2-\hat{\alpha}_2)\mathbf{K}_{12}^r] - \\ &(\alpha_1-\hat{\alpha}_1)\sum_{j\neq 1,2}^{N_r}(\alpha_j-\hat{\alpha}_j)\mathbf{K}_{1j}^r - (\alpha_2-\hat{\alpha}_2)\sum_{j\neq 1,2}^{N_r}(\alpha_j-\hat{\alpha}_j)\mathbf{K}_{2j}^r + Re(y_1)(\alpha_1-\hat{\alpha}_1) + Re(y_2)(\alpha_2-\hat{\alpha}_2) \\ &-\epsilon(\alpha_1+\hat{\alpha}_1+\alpha_2+\hat{\alpha}_2),\end{aligned} \tag{63}$$

Based on (63), the partial derivative of $\Theta^r_{1,2}$ with respect to α_1 is

$$\frac{\partial \Theta^r_{1,2}}{\partial \alpha_1} = -(\alpha_1 - \hat{\alpha}_1)\mathbf{K}^r_{11} - (\alpha_2 - \hat{\alpha}_2)\mathbf{K}^r_{12} - \sum_{j \neq 1,2}^{N_r} (\alpha_j - \hat{\alpha}_j)\mathbf{K}^r_{1j} + Re(y_1) - \epsilon =$$

$$-(\alpha_1 - \hat{\alpha}_1)^{new}\mathbf{K}^r_{11} + Re(y_1) - \sum_{j \neq 1}^{N_r} (\alpha_j - \hat{\alpha}_j)\mathbf{K}^r_{1j} - \epsilon = -(\alpha_1 - \hat{\alpha}_1)^{new}\mathbf{K}^r_{11} + \Phi_1 + (\alpha_1 - \hat{\alpha}_1)\mathbf{K}^r_{11} - \epsilon$$

$$\Rightarrow \alpha_1^{new} = \alpha_1 + \frac{(\Phi^r_1 - \epsilon)}{\mathbf{K}^r_{11}}, \tag{64}$$

Similarly we can derive the update formulas of $\hat{\alpha}_1$,α_2 and $\hat{\alpha}_2$

$$\hat{\alpha}_1^{new} = \hat{\alpha}_1 - \frac{(\Phi^r_1 + \epsilon)}{\mathbf{K}^r_{11}}, \tag{65}$$

$$\alpha_2^{new} = \alpha_2 + \frac{(\Phi^r_2 - \epsilon)}{\mathbf{K}^r_{22}}, \tag{66}$$

$$\hat{\alpha}_2^{new} = \hat{\alpha}_2 - \frac{(\Phi^r_2 + \epsilon)}{\mathbf{K}^r_{22}}, \tag{67}$$

It is obviously the dual variables in 2-D solver have the same update rule as that of 1-D solver. Based on (53), assume the ith and jth dual variable pair are chosen, the gain of 2-D solver objective function can be written as

$$\bigtriangledown \Theta^r_{ij} = -\frac{1}{2}[(\delta_i - \hat{\delta}_i)^2\mathbf{K}^r_{ii} + (\delta_j - \hat{\delta}_j)^2\mathbf{K}^r_{jj} + 2(\delta_i - \hat{\delta}_i)(\delta_j - \hat{\delta}_j)\mathbf{K}^r_{ij}] + \Phi^r_i(\delta_i - \hat{\delta}_i) + \Phi^r_j(\delta_j - \hat{\delta}_j)$$

$$-\epsilon(\delta_i + \hat{\delta}_i + \delta_j + \hat{\delta}_j), \tag{68}$$

recall the gain of objective function of 1-D solver in (61), we obtain

$$\bigtriangledown\Theta_{ij}^r = \bigtriangledown\Theta_i^r + \bigtriangledown\Theta_j^r - (\delta_i - \hat{\delta}_i)(\delta_j - \hat{\delta}_j)\mathbf{K}_{ij}^r, \tag{69}$$

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

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

From (69), it is obviously that the gain of 2-D solver is a summation of the gain of 2 independent 1-D solver and a correlation term $(\delta_i + \hat{\delta}_i)(\delta_j + \hat{\delta}_j)\mathbf{K}_{ij}^r$.

Obviously optimal 2-D coordinate combination (i, j) can be determined by comparing the gains of all the possibilities exhaustively, which requires $O(n^2)$ times of searching. Based on (69), we can approximate optimal 2-D solution by 1-D search approach, we will prove in large MIMO scenario, when N_t is sufficient large, with channel hardening become effective, this approximation is very efficient. Here we propose two kinds of 1-D approximate searching strategy:

1. 1-D searching without damping:

do 1 time 1-D searching and calculate all the 1-D gain based on (61), then choose the coordinate pairs with first and second largest 1-D gain as the candidate.

2. 1-D searching with damping:

do 2 times 1-D searching, in the first round find dual variable pair i that can maximize 1-D gain, then in the second round, find jth dual variable pair with the value of ith coordinate updated.

From (69), it can be easily interpreted the efficient of 1-D approximation approach is majorly determined by the approximation ratio $\frac{(\delta_i+\hat{\delta}_i)(\delta_j+\hat{\delta}_j)\mathbf{K}_{ij}^r}{\bigtriangledown\Theta_i^r+\bigtriangledown\Theta_j^r}$, hence we provide theoretical analyse from the view of channel hardening phenomenon, and give the upper bound of approximation ratio. Prior the theoretical analyse, we first investigate some mathematical properties of channel hardening.

Based on the complimentary KKT conditions, we have $\alpha\hat{\alpha} = 0$, therefore we transform (49) and (53) by substituting $\lambda_i = \alpha_i - \hat{\alpha}_i$ and $|\lambda_i| = \alpha_i + \hat{\alpha}_i$, hence we have

$$maximize \quad \Theta_S^r = -\frac{1}{2}\lambda_S^T \mathbf{K}_{SS}^r \lambda_S + (Re(\mathbf{y})_S^T - \lambda_N^T \mathbf{K}_{NS}^r)\lambda_S - \epsilon(\mathbf{e}_S^T |\lambda_s|), \tag{70}$$

$$\bigtriangledown\Theta_S^r = -\frac{1}{2}\tau_S^T \mathbf{K}_{SS}^r \tau_S + (\Phi_S^r)^T \tau_S - \epsilon \mathbf{e}_S^T (|\lambda|^{new} - |\lambda|)_S, \tag{71}$$

where $\tau = \lambda^{new} - \lambda$. Therefore we can simplify the work set selection procedure by update variables λ rather than dual variable pairs For single direction solver, (54) can be transformed as

$$maximize \quad \Theta_1^r = -\frac{1}{2}(\lambda_1^{new})^2 \mathbf{K}_{11}^r - (\lambda_1)^{new} \sum_{j=2}^{N_r} \mathbf{K}_{1j}^r \lambda_j + Re(y_1)\lambda_1^{new} - \epsilon|\lambda_1|^{new}, \tag{72}$$

Thus for 1-D solver the gradient of (72) with respect to λ can be written as

$$\lambda_1^{new} = \lambda_1 + \frac{\Phi_1 - sgn(\lambda_1^{new})\epsilon}{\mathbf{K}_{11}^r}, \tag{73}$$

where $\frac{\partial|\lambda_1^{new}|}{\partial\lambda_1^{new}} = sgn(\lambda_1^{new})$ the similar principle can be also applied to double direction solver. Because in the update process the $sgn(\lambda^{new})$ is unknown to old parameter values , therefore, we need to consider both the case $sgn(\lambda^{new}) = -1$ or 1 and choose the one with the larger objective function gain $\bigtriangledown\Theta_S^r$. Here we give the pseudo code for sequential approximate optimal double direction solver.

Algorithm 1 Dual Channel Complex Support Vector Detection Algorithm

```
procedure CSVD(y,H)
    Step 1. Initialization
    for $i = 1 : N_r$ do        ▷ initialize $\lambda^r$, $\lambda^i$, $\Phi^r$, $\Phi^i$, $\Psi^r$, $\Psi^i$ and duality gap $G$
        $\lambda_i^r = 0, \lambda_i^i = 0$
        $\Phi_i^r = Re(y_i), \Phi_i^i = Im(y_i)$
        $\Psi_i^r = 0, \Psi_i^i = 0$
    end for
    Step 2. if $G > tol$, go to step 3, else go to Step 6
    Step 3.
    Sequential 2-D solver with or without damping        ▷ find two dual variables to be updated
    Step 4.
    $G$
    Step 5.
    $\tilde{x} = (\lambda^r + i\lambda^i)\mathbf{H}$        ▷ reconstruct $\mathbf{x}$
    $\mathbf{x} = \mathbb{Q}(\tilde{x})$        ▷ $\mathbb{Q}(\cdot)$ denotes quantization operation based on symbol constellation
    go back to Step 2
    Step 6. Return $\mathbf{x}$
end procedure
```

Algorithm 2 Sequential 2-D Solver without Damping

```
procedure SEQUENTIAL 2-D SOLVER WITHOUT DAMPING($1st$, $2nd$)
    for $i = 1 : N_r$ do
        update $\lambda_i^r(\lambda_i^i)$ by single direction solver
        calculate $\bigtriangledown\theta_i^r(\bigtriangledown\theta_i^i)$        ▷ calculate the gain of sub objective function
    end for
    choose the dual variable with first and the second largest gain of sub objective function,
denoted as 1st and 2nd
    update $\Phi^r(\Phi^i)$ and $\Psi^r(\Psi^i)$ with respect to 1st and 2nd
    Return 1st, 2nd
end procedure
```

VI. INITIALIZATION

Computer simulations are made for different sizes of V-BLAST MIMO systems, with $5 \leq N_r \leq 100, 5 \leq N_t \leq N_r$, the empirical estimation of logarithmic expectation of ϕ_{om}, $E[\ln(\phi_{om})]_{em}$,

Algorithm 3 Sequential 2-D Solver with Damping

```
procedure SEQUENTIAL 2-D SOLVER WITH DAMPING($1s_1t$,$1st_2$)
    for $i = 1 : N_r$ do                                        ▷ First round searching
        update $\lambda_i^r(\lambda_i^i)$ by single direction solver
        calculate $\bigtriangledown\theta_i^r(\bigtriangledown\theta_i^i)$        ▷ calculate the gain of objective function
    end for
    choose the dual 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
        update $\lambda_i^r(\lambda_i^i)$ by single direction solver
        calculate $\bigtriangledown\theta_i^r(\bigtriangledown\theta_i^i)$        ▷ calculate the gain of objective function
    end for
    choose the dual variable with the largest gain of objective function as $1st_2$
    update $\Phi^r(\Phi^i)$ and $\Psi^r(\Psi^i)$ with respect to $1st_2$
    Return $1st_1$ and $1st_2$
end procedure
```

is calculated by taking average over $1e4$ channel realizations for each size of MIMO systems, as shown in Fig.**??**, the Theoretical logarithmic expectation of ϕ_{om} $E[\ln(\phi_{om})]_t$ in (**??**) is plotted in Fig.**??**. Average deviation between $E[\ln(\phi_{om})]_{em}$ and $E[\ln(\phi_{om})]_t$ is also calculated, $V_{em-t} = 7.3043e - 04$.

Fig.**??** demonstrates the relation between the number of users (N_t) and $E[\ln(\phi_{om})]_t$ under cases of different numbers of antennas at base station (N_r). From Fig.**??**, we can see, on the one hand, with N_r fixed, $E[\ln(\phi_{om})]$ decreases while N_t increases, however the gradient of each curve becomes more and more gentle. On the other hand, when N_r becomes larger $E[\ln(\phi_{om})]$ becomes more insensitive to variation of N_t.

VII. Stopping Criteria

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

$$\theta(\lambda^r, \lambda^i) = f(\mathbf{w}, \xi) \tag{74}$$

$$\frac{1}{2}||\mathbf{w}||_{\mathbb{H}}^2 + C\sum_{i=1}^{N_r}[R(\xi_i^r) + R(\hat{\xi}_i^r) + R(\xi_i^i) + R(\hat{\xi}_i^i)], \tag{75}$$

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

$$\begin{aligned}\theta(\lambda^r, \lambda^i) = -\frac{1}{2} < (\lambda^r)^T, \mathbf{K}^r\lambda^r > -\frac{1}{2} < (\lambda^i)^T, \mathbf{K}^r\lambda^i > + < Re(\mathbf{y})^T, \lambda^r > + < Im(\mathbf{y})^T, \lambda^i >\\ -\epsilon < \mathbf{e}^T, (|\lambda^r| + |\lambda^i|) > +C\sum_{i=1}^{N_r}[\tilde{R}(\xi_i^r) + \tilde{R}(\hat{\xi}_i^r) + \tilde{R}(\xi_i^i) + \tilde{R}(\hat{\xi}_i^i)],\end{aligned} \tag{76}$$

Similarly, (39) can be formulated as

$$||\mathbf{W}||_{\mathbb{H}}^2 =< (\lambda^r)^T, \mathbf{K}^r\lambda^r > + < (\lambda^i)^T, \mathbf{K}^r\lambda^i > -2 < \lambda^r, \mathbf{K}^i\lambda^i >, \tag{77}$$

hence, duality gap can be formulated as

$$\begin{aligned}G(\lambda^r, \lambda^i) = f(\mathbf{w}, \xi) - \theta(\lambda^r, \lambda^i) =< (\lambda^r)^T, \mathbf{K}^r\lambda^r > + < (\lambda^i)^T, \mathbf{K}^r\lambda^i > - < Re(\mathbf{y})^T, \lambda^r > - < Im(\mathbf{y})^T, \lambda^i >\\ -\epsilon < \mathbf{e}^T, (|\lambda^r| + |\lambda^i|) > +C\sum_{i=1}^{N_r}[\xi_i^r R^{'}(\xi_i^r) + \hat{\xi}_i^r R^{'}(\hat{\xi}_i^r) + \xi_i^i R^{'}(\xi_i^i) + \hat{\xi}_i^i R^{'}(\hat{\xi}_i^i)] - 2 < \lambda^r, \mathbf{K}^i\lambda^i > .\end{aligned} \tag{78}$$

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

$$R(\xi) = \frac{1}{2}\xi^2, \tag{79}$$

hence

$$\tilde{R}(\xi) = R(\xi) - \xi R'(\xi) = -\frac{1}{2}\xi^2, \tag{80}$$

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

$$\xi_i^r = Re(\mathbf{y}_i) - Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - \epsilon \tag{81}$$

$$\hat{\xi}_i^r = Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}}) - Re(\mathbf{y}_i) - \epsilon \tag{82}$$

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

$$\xi_i^r(\hat{\xi}_i^r) = \max(0, |Re(\mathbf{y}_i) - Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})| - \epsilon) \tag{83}$$

$$\xi_i^i(\hat{\xi}_i^i) = \max(0, |Im(\mathbf{y}_i) - Im(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})| - \epsilon) \tag{84}$$

$$\xi\hat{\xi} = 0, \tag{85}$$

Therefore the risk function can be rewritten as

$$R(\xi_i^r) + R(\hat{\xi}_i^r) = \frac{1}{2}(|Re(\mathbf{y}_i) - Re(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})|)_\epsilon^2 \tag{86}$$

$$R(\xi_i^i) + R(\hat{\xi}_i^i) = \frac{1}{2}(|Im(\mathbf{y}_i) - Im(< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}})|)_\epsilon^2 \tag{87}$$

where $()_\epsilon$ denotes ϵ insensitive function as we mention in section III-B. Based on (38), we have

$$Re(\mathbf{y}_i) - Re(<\mathbf{h}_i, \mathbf{W}>_{\mathbb{H}}) = Re(\mathbf{y}_i) - \sum_{j=1}^{N_r} \lambda_j^r \mathbf{K}_{ij}^r + \sum_{j=1}^{N_r} \lambda_j^i \mathbf{K}_{ij}^i \tag{88}$$

$$Im(\mathbf{y}_i) - Im(<\mathbf{h}_i, \mathbf{W}>_{\mathbb{H}}) = Im(\mathbf{y}_i) - \sum_{j=1}^{N_r} \lambda_j^i \mathbf{K}_{ij}^r - \sum_{j=1}^{N_r} \lambda_j^r \mathbf{K}_{ij}^i \tag{89}$$

we define two intermediate variables Φ and Ψ

$$\Phi^r = Re(\mathbf{y}) - \mathbf{K}^r \lambda^r; \Phi^i = Im(\mathbf{y}) - \mathbf{K}^r \lambda^i \tag{90}$$

$$\Psi^r = \mathbf{K}^i \lambda^i; \Psi^i = -\mathbf{K}^i \lambda^r \tag{91}$$

Therefore based on (87)-(92), duality gap in (78) can be rewritten as

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

$$-\epsilon < \mathbf{e}^T, (|\lambda^r| + |\lambda^i|) > +C \sum_{i=1}^{N_r} [(|\Phi_i^r + \Psi_i^r|)_\epsilon^2 + (|\Phi_i^i + \Psi_i^i|)_\epsilon^2] - 2 < \lambda^r, \mathbf{K}^i \lambda^i > . \tag{92}$$

The duality gap between primal problem and dual problem is used to evaluate how close a solution is to global minimum. In our scenario, duality gap is employed as stopping criteria because of low computational cost and reliable performance comparing to monitor KKT complementary condition.

VIII. Hyperparameter Setting

IX. Computer Simulations

X. Conclusion

The conclusion goes here.

Appendix one text goes here.

Appendix A

Appendix two text goes here.

Appendix B

Let $\mathbf{A} \in \mathbb{C}^{m\times m}$, $A \sim \mathbb{C}W(n, \mathbf{\Sigma})$, $\mathbb{C}W(n, \mathbf{\Sigma})$ denotes complex Wishart distribution with n degrees of freedom and covariance matrix $\mathbf{\Sigma}$. It is obvious $\mathbf{A}$ is Hermition positive definite matrix, $\mathbf{A} = \mathbf{A}^H > 0$.

The pdf of $\mathbf{A}$ can be written as [**?**]:

$$f(\mathbf{A}) = \{\tilde{\Gamma}_m(n) det(\mathbf{\Sigma})^n\}^{-1} det(\mathbf{A})^{n-m} etr(-\mathbf{\Sigma}^{-1}\mathbf{A}), \tag{93}$$

where $\tilde{\Gamma}_m(\beta)$ denotes multivariate complex Gamma function defined by:

$$\tilde{\Gamma}_m(\beta) = \pi^{\frac{m(m-1)}{2}} \prod_{i=1}^{m} \Gamma(\beta - i + 1) \quad Re(\beta) > m - 1. \tag{94}$$

Furthermore, from [**?**], we have

$$\tilde{\Gamma}_m(\beta) = \int_{\mathbf{X}=\mathbf{X}^H>0} etr(-\mathbf{X}) det(\mathbf{X})^{\beta-m} d\mathbf{X} \quad Re(\beta) > m - 1. \tag{95}$$

We derive logarithmic expectation of $det(\mathbf{A})$

$$\begin{aligned} E[\ln(det(\mathbf{A}))] &= \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))f(\mathbf{A})d\mathbf{A} \\ &= \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))\{\tilde{\Gamma}_m(n)det(\mathbf{\Sigma})^n\}^{-1}det(\mathbf{A})^{n-m}etr(-\mathbf{\Sigma}^{-1}\mathbf{A})d\mathbf{A} \\ &= \frac{det(\mathbf{\Sigma})^{-n}}{\tilde{\Gamma}_m(n)} \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))det(\mathbf{A})^{n-m}etr(-\mathbf{\Sigma}^{-1}\mathbf{A})d\mathbf{A}, \qquad (96) \end{aligned}$$

if $\mathbf{\Sigma}=\mathbf{I}$, (96) can be written as

$$E[\ln(det(\mathbf{A}))] = \frac{1}{\tilde{\Gamma}_m(n)} \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))det(\mathbf{A})^{n-m}etr(-\mathbf{A})d\mathbf{A}. \qquad (97)$$

Because $\frac{d}{dn}[det(\mathbf{A})]^{n-m} = \ln(det(\mathbf{A}))det(\mathbf{A})^{n-m}$, (97) can be rewritten as

$$E[\ln(det(\mathbf{A}))] = \frac{1}{\tilde{\Gamma}_m(n)} \frac{d}{dn} \int_{\mathbf{A}=\mathbf{A}^H>0} etr(-\mathbf{A})det(\mathbf{A})^{n-m}d\mathbf{A}, \qquad (98)$$

using (95), (98) can be rewritten as

$$E[\ln(\mathbf{A})] = \frac{\tilde{\Gamma}'_m(n)}{\tilde{\Gamma}_m(n)}. \qquad (99)$$

Based on (94), we have

$$\tilde{\Gamma}'_m(n) = \pi^{\frac{m(m-1)}{2}} \sum_{i=1}^{m} [\Gamma'(n-i+1) \prod_{j=1, j\neq i}^{m} \Gamma(n-j+1)], \qquad (100)$$

Thus we have

$$E[\ln(det(\mathbf{A}))] = \frac{\tilde{\Gamma}'_m(n)}{\tilde{\Gamma}_m(n)} = \sum_{i=1}^{m} \frac{\Gamma'(n-i+1)}{\Gamma(n-i+1)} = \sum_{i=1}^{m} \psi(n-i+1), \qquad (101)$$

where ψ denotes Digamma function.

APPENDIX C

If $x \sim Gamma(n, \theta)$, with shape parameter k and scale parameter θ, $x > 0$, $\Gamma(k)$ denotes Gamma function, the density function of Gamma distribution is

$$f(x, k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k}. \tag{102}$$

Thus we have

$$E[\ln(x)] = \frac{1}{\Gamma(k)} \int_0^\infty \ln(x) x^{k-1} e^{-x/\theta} \theta^{-k} dx, \tag{103}$$

define $z = x/\theta$ and since $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$, (103) can be rewritten as

$$E[\ln(x)] = \ln(\theta) + \frac{1}{\Gamma(k)} \int_0^\infty \ln(z) z^{k-1} e^{-z} dz. \tag{104}$$

Because $\frac{d(z^{k-1})}{dk} = \ln(z) z^{k-1}$, (104) can be rewritten as

$$\begin{aligned} E[\ln(z)] &= \ln(\theta) + \frac{1}{\Gamma(k)} \frac{d}{dk} \int_0^\infty z^{k-1} e^{-z} dz \\ &= \ln(\theta) + \frac{\Gamma'(k)}{\Gamma(k)} \\ &= \ln(\theta) + \psi(k), \end{aligned}$$

where $\psi(k)$ denotes Digamma function.

APPENDIX D

$x_1, x_2, \cdots x_{N_t}$ are independent beta variables, the probability density function (pdf) can be written as:

$$f(x_i) = \frac{1}{\mathbb{B}(k_1^i, k_2^i)} x_i^{k_1^i - 1} (1 - x_i)^{k_2^i - 1}, \tag{105}$$

define $y_i = -\ln(x_i) = g(x_i)$, Based on Jacobian transformation, we have

$$f_{y_i}(\rho) = |\frac{dy_i}{dx_i}|^{-1} f_{x_i}(g^{-1}(\rho)) = \frac{1}{\mathbb{B}(k_1^i, k_2^i)} e^{-k_1^i \rho}(1 - e^{-\rho})^{k_2^i - 1}. \tag{106}$$

where (106) can be alternatively expressed as [**?**]

$$f_{y_i}(\rho) = \sum_{j^i=0}^{k_2^i - 1} c(k_1^i, k_2^i, j^i)(k_1^i + j^i) exp(-(k_1^i + j^i)\rho), \tag{107}$$

where $c(k_1^i, k_2^i, j_i) = (-1)^{j^i} \binom{k_2^i - 1}{j^i} [(k_1^i + k_2^i)\mathbb{B}(k_1^i, k_2^i)]^{-1}$, $\mathbb{B}(\alpha, \beta)$ denotes beta function. Based on the lemma 1 of [**?**], if $a_1, a_2, \cdots a_n$ are independent exponentially distributed random variables, with pdf given by

$$t_i exp(-t_i a_i) \tag{108}$$

then pdf of $a = \sum_{i=1}^n a_i$ can be written as

$$f(a|\mathbf{t}) = \prod_{i=1}^n t_i \sum_{i=1}^n [exp(-t_i a) / \prod_{j=1 j\neq i}^{j=n} (t_j - t_i)], \tag{109}$$

where $t = [t_1, t_2, \cdots t_n]$. The pdf of y_i can be viewed as the weighting summation of exponential distribution functions, define $y = \sum_{i=1}^n y_i$, based on (109), the pdf of y is given by

$$f_y(m) = \sum_{\mathbf{j}} \{[\prod_{i=1}^n c(k_1^i, k_2^i, j^i)] f(m|\mathbf{k_1} + \mathbf{j})\}, \tag{110}$$

where $\sum_{\mathbf{j}} = \sum_{j^1} \sum_{j^2} \cdots \sum_{j^n}$, the range of j^i is defined by $j^i \in [0, k_2^i]$, $f(m|\mathbf{k_1} + \mathbf{j}) = (\prod_{i=1}^{N_t}(k_1^i + j^i)) \sum_{i=1}^{N_t} [exp(-(k_1^i + j^i)m) / \prod_{j=1 j\neq i}^{N_t}(k_1^j + j^j - k_1^i - j^i)]$, $\mathbf{k_1} + \mathbf{j} = [k_1^1 + j^1, k_1^2 + j^2 \cdots k_1^n + j^n]$. we define $U = exp(-y) = \prod_{i=1}^n x_i$, using Jacobian transformation, the pdf of U

is given by

$$f_U(u) = |\frac{du}{dy}|^{-1} f_y(-\ln(u)) = \frac{1}{u} \sum_{\mathbf{j}} \{ [\prod_{i=1}^{n} c(k_1^i, k_2^i, j^i)] f(-\ln(u)|\mathbf{k_1} + \mathbf{j}) \}. \tag{111}$$

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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Hoydis, S. Ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 2, pp. 160–171, 2013.

[11] 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.

[12] 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.

[13] Y. Sun, "A family of likelihood ascent search multiuser detectors," *submitted to IEEE Trans. on Information Theory*.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[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. 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.

[20] 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.

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

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

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

[24] 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.

[25] 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.

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

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

[28] B. Schölkopf, C. Burges, and V. Vapnik, “Incorporating invariances in support vector learning machines,” in *Artificial Neural NetworksICANN 96*. Springer, 1996, pp. 47–52.

[29] 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.

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

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

[32] 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.

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

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