Report

Tianpei Chen

Department of Electrical and Computer Engineering

McGill University

August 20, 2015

I. Introduction

Decoder is one of the key components of Multiple-Input Multiple-Output (MIMO) systems. Designing of high performance and low complexity detector has become a bottleneck of Large MIMO systems.

Firmly grounded in framework of statistical learning theory, Support Vector Machine (SVM) is proposed in 1960s [ref vapnik], and of immense research and industry interest since 1990s. SVM is a powerful tool for supervised learning tasks such as classification, regression and prediction. Moreover, the kernel trick [ref learning with kernels SVM regularization] makes it possible to map data samples into higher dimensional feature space. Therefore SVM can deal with non-linear learning tasks. This makes SVM become a promising tool for complex real-world problems. Based on the similar principle, ϵ-Support Vector Regression (epsilon-SVR) [vapnik 1995, smola 2003], is developed.

Like SVM, epsilon-SVR first change primal objective function into dual optimization task, then solving the dual quadratic optimization problem. Typically this kind of problem can be solved by numerical quadratic optimization (QP) methods, however, they are computational costly. Decomposition methods, denotes a set of algorithms that divide the optimization variables (Lagrange multipliers) into two sets W and N, W is the work set and N contains the rest optimization variables. In each iteration, only work set is updated for optimization while the other variables are fixed. Sequential Minimal Optimization (SMO) [ref A fast algorithm sequential minimal optimization] is an extreme case of decomposition methods which chooses dual Lagrange multiplier to optimize in each iteration. In each iteration, decomposition method can find an analytic optimal solution for work set, which makes the solver works much more faster than numerical QP algorithms. Decomposition methods can be employed to epsilon SVR by the similar manner.

Bouloulis et employ Wirtingers calculus into Reproducing Kernel Hilbert Space (RKHS) so that expands real-SVM to pure complex SVM by exploiting complex kernel [ref complex support vector machine]. 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 be divided into two parallel real SVR optimization problem 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.

Steinwart et[ref SVM without offset] shows with a proper designed work set selection strategy, the approach that choosing double Lagrange multipliers can be much more faster than choosing single Lagrange multiplier without performance loss.

Based on the discrete time MIMO channel model, In our regression model, this CSVR-detector is constructed without offset, The offset in SVR imposes an additional linear quality constraint, which makes it necessary for decomposition methods such as Sequential Minimal Optimization to update more than one Lagrange multipliers in each iteration.

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 Lagrange multiplier search strategy that find two Lagrange multiplier sequentially, which can approximate the optimal dual Lagrange multiplier searching strategy. The former one only requires $O(n)$ searches in one iteration, while the optimal dual Lagrange multiplier strategy requires $O(n^2)$ searches per iteration.

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.

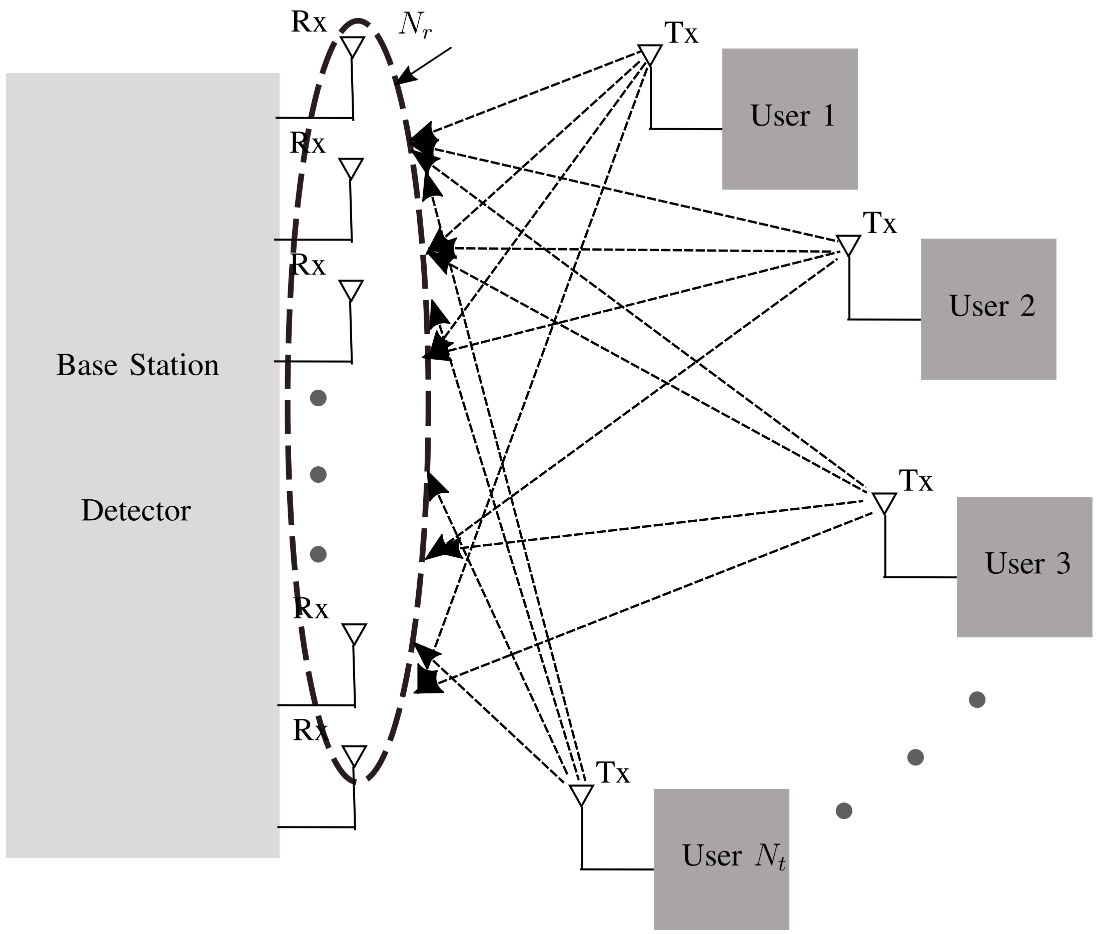

Fig. 1: Large MIMO uplink system

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:

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

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

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, 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 cost function, the simplest cost 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.

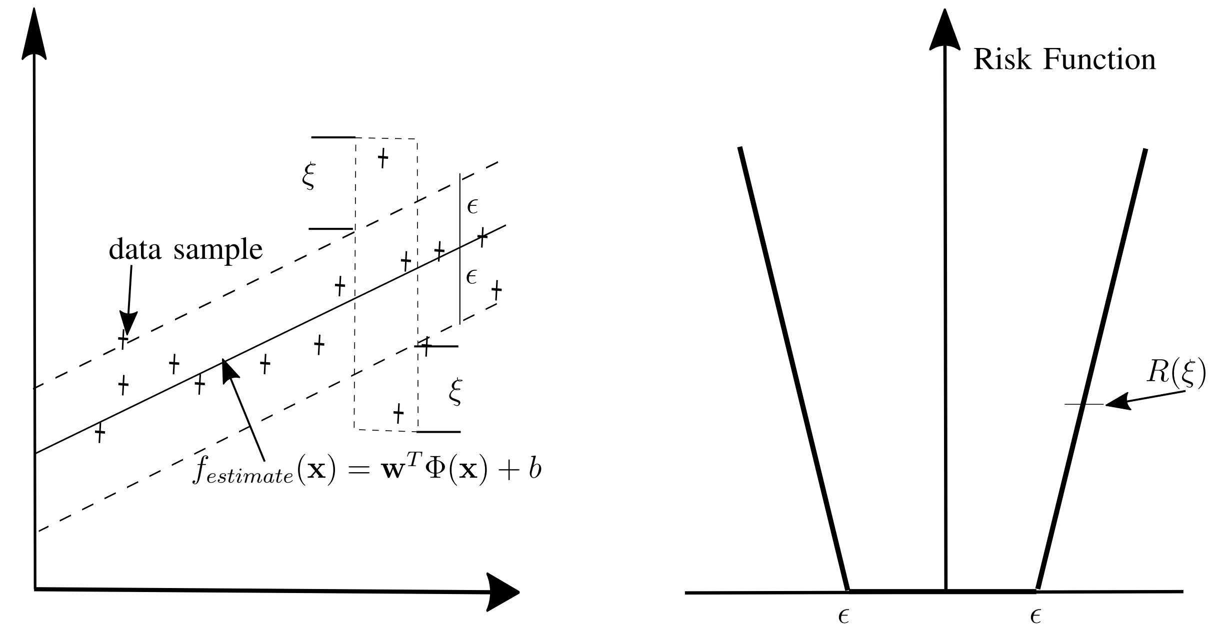

Fig. 2

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

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

Take the logarithm of (4), 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

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

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

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

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):

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

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 [**?**].

we construct dual form of (3) by introducing Lagrange multiplier, we have dual optimization problem

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

where η,$\hat{\eta}$, α, $\hat{\alpha}$ are Lagrange multipliers, the solution of dual optimization problem is saddle point, from [**?**], the dual problem of (13) can be formulated as

$$
\begin{aligned}
\max_{\alpha,\hat{\alpha},\eta,\hat{\eta}} \min_{\mathbf{w},\xi,\hat{\xi}} \Theta = \frac{1}{2}||\mathbf{w}||^2 + \sum_{j=1}^{l} C_i(R(\xi_i) + R(\hat{\xi}_i)) - \sum_{i=1}^{l}(\eta_i\xi_i + \hat{\eta}_i\hat{\xi}_i) \\
-\sum_{i=1}^{l} \alpha_i(\epsilon + \xi_i - y_i + \mathbf{w}^T\Phi(\mathbf{x}_i)) - \sum_{i=1}^{l} \hat{\alpha}_i(\epsilon + \hat{\xi}_i + y_i - \mathbf{w}^T\Phi(\mathbf{x}_i)) \\
s.t. \begin{cases} \eta, \hat{\eta}, \alpha, \hat{\alpha} \geq 0 \\ \xi, \hat{\xi} \geq 0 \end{cases}
\end{aligned} \tag{14}
$$

KarushKuhnTucker (KKT) condition is satisfied or not [] determines whether a non linear optimization problem find its optimal solution, the complementary KKT solution of (14) is

$$\begin{cases} \eta_i \xi_i = 0 \\ \hat{\eta}_i \hat{\xi}_i = 0 \\ \alpha_i(\epsilon + \xi_i - y_i + \mathbf{w}^T \Phi(\mathbf{x}_i)) = 0 \\ \hat{\alpha}_i(\epsilon + \hat{\xi}_i + y_i - \mathbf{w}^T \Phi(\mathbf{x}_i)) = 0 \end{cases} \tag{15}$$

Take the partial derivative of Θ with respect to $\mathbf{w}$, ξ, $\hat{\xi}$ and b, and find the minimum.

$$\frac{\partial \theta}{\partial \mathbf{w}} = w - \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i) \tag{16}$$

$$\frac{\partial \theta}{\partial \xi} = C_i R^{'}(\xi_i) - \eta_i - \alpha_i = 0 \tag{17}$$

$$\frac{\partial \theta}{\partial \hat{\xi}} = C_i R^{'}(\hat{\xi}_i) - \hat{\eta}_i - \hat{\alpha}_i = 0 \tag{18}$$

$$\frac{\partial \theta}{\partial b} = \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i) = 0 \tag{19}$$

Then substitute (16)-(19) to (14), yields the final dual optimization objective function, for sake of brevity, we make C_i uniform to all data samples

$$\begin{aligned} maximize \quad \Theta \quad &= \quad \frac{1}{2} \sum_{i=1}^{l} \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j) \Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_i) + C \sum_{i=1}^{l} [(R(\xi_i) - \xi_i R^{'}(\xi_i)) \\ &+ \quad (R(\hat{\xi}_i) - \hat{\xi}_i R^{'}(\hat{\xi}_i))] + \sum_{i=1}^{l} [(\alpha_i - \hat{\alpha}_i) y_i - (\alpha_i + \hat{\alpha}_i) \epsilon] \\ &- \quad \sum_{i=1}^{l} \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j) \Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_i), \\ &s.t. \quad \begin{cases} \sum_{i=1}^{l} (\alpha_i - \hat{\alpha}_i) = 0 \\ 0 < \alpha < C \tilde{R}^{'}(\alpha) \\ 0 < \hat{\alpha} < C \tilde{R}^{'}(\hat{\alpha}) \end{cases} \end{aligned} \tag{20}$$

(20) can be simplified to

$$
\begin{aligned}
maximize \quad \Theta &= -\frac{1}{2}\sum_{i=1}^{l}\sum_{i=1}^{l}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\Phi(\mathbf{x}_j)^T\Phi(\mathbf{x}_i)+\sum_{i=i}^{l}[(\alpha_i-\hat{\alpha}_i)y_i-(\alpha_i+\hat{\alpha}_i)\epsilon] \\
&+ C\sum_{i=1}^{l}[\tilde{R}(\xi_i)+\tilde{R}(\hat{\xi}_i)] \\
&= -\frac{1}{2}(\mathbf{a}-\hat{\mathbf{a}})^T\mathbf{K}(\mathbf{a}-\hat{\mathbf{a}})+(\mathbf{y}-\epsilon)^T\mathbf{a}+(-\mathbf{y}-\epsilon)^T\hat{\mathbf{a}}+\mathbf{e}^TC(\tilde{R}(\xi)+\tilde{R}(\hat{\xi})), \quad (21)
\end{aligned}
$$

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 $\mathbf{a}^{(*)}$= $\left[\begin{smallmatrix}\mathbf{a}\\ \hat{\mathbf{a}}\end{smallmatrix}\right]$, $\mathbf{v}\in\mathbb{R}^{2l}$,

$$
\mathbf{v}_i=\begin{cases}1 & i=1,\ldots,l\\ -1 & i=l+1,\ldots,2l\end{cases} \tag{22}
$$

(21) can be reformulate as

$$
maximize \quad \Theta=-\frac{1}{2}(\mathbf{a}^*)^T\left[\begin{smallmatrix}\mathbf{K} & -\mathbf{K}\\ -\mathbf{K} & \mathbf{K}\end{smallmatrix}\right]\mathbf{a}^{(*)}+[(\mathbf{y}-\epsilon)^T,(-\mathbf{y}-\epsilon)^T]\mathbf{a}^{(*)}+\mathbf{e}^TC(\tilde{R}(\xi)+\tilde{R}(\hat{\xi})), \tag{23}
$$

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 $(\mathbf{h}_1,y_1)(\mathbf{h}_2,y_2),\ldots,(\mathbf{h}_{N_r},y_{N_r})$, where $\mathbf{h}_i$ denotes ith row of channel propagation matrix $\mathbf{H}$, this yields a regression task without offset b:

$$
y_i=f_{true}(\mathbf{h}_i)+n, \tag{24}
$$

$$
f_{true}(\mathbf{h}_i)=\mathbf{h}_i\mathbf{s}, \tag{25}
$$

$$
\tag{26}
$$

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

$$\mathbf{a}^T\mathbf{b}\in<\mathbf{a},\mathbf{b}>_H, \tag{27}$$

From Property 1 and Property 3, it is obvious

$$<\mathbf{g},\mathbf{h}>_{\mathbb{H}}=<\mathbf{g}^r,\mathbf{h}^r>_H+<\mathbf{g}^i,\mathbf{h}^i>_H+i(<\mathbf{g}^i,\mathbf{h}^r>_H-<\mathbf{g}^r,\mathbf{h}^i>_H) \tag{28}$$

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

$$\begin{aligned}<\mathbf{h},\mathbf{s}^*>_{\mathbb{H}} &= <\mathbf{h}^r,(\mathbf{s}^*)^r>_H+<\mathbf{h}^i,(\mathbf{s}^*)^i>_H+i(<\mathbf{h}^i,(\mathbf{s}^*)^r>_H-<\mathbf{h}^r,(\mathbf{s}^*)^i>_H)\\ &= <\mathbf{h}^r,\mathbf{s}^r>_H-<\mathbf{h}^i,\mathbf{s}^i>_H+i(<\mathbf{h}^i,\mathbf{s}^r>_H+<\mathbf{h}^r,\mathbf{s}^i>_H),\end{aligned} \tag{29}$$

$$\mathbf{h}\mathbf{s}=\mathbf{h}^r\mathbf{s}^r-\mathbf{h}^i\mathbf{s}^i+i(\mathbf{h}^i\mathbf{s}^r+\mathbf{h}^r\mathbf{s}^i), \tag{30}$$

Because of (27), (29) and (30), $\mathbf{h}_i\mathbf{s} \in < \mathbf{h}_i, \mathbf{s}^* >_{\mathbb{H}}$ $\square$

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

$$minimize \quad \frac{1}{2}||w||^2_{\mathbb{H}} + C\sum_{k=1}^{N_r}[(R(\xi^r_k) + R(\hat{\xi}^r_k) + R(\xi^i_k) + R(\hat{\xi}^i_k))]$$
$$s.t. \begin{cases} Re(y_k - < \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) \leq \epsilon + \xi^r_k \\ Re(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}} - y_k) \leq \epsilon + \hat{\xi}^r_k \\ Im(y_k - < \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}) \leq \epsilon + \xi^i_k \\ Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}} - y_k) \leq \epsilon + \hat{\xi}^i_k \\ \xi^r, \hat{\xi}^r, \xi^i, \hat{\xi}^i \geq 0 \end{cases} \tag{31}$$

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 (14), Lagrange duality is formulated for (31)

$$\max_{(\alpha,\hat{\alpha},\beta,\hat{\beta},\eta,\hat{\eta},\tau,\hat{\tau})} \min_{(\mathbf{w},\xi^r,\hat{\xi}^r,\hat{\xi}^i,\hat{\xi}^i)} \theta = \frac{1}{2}||w||^2_{\mathbb{H}} + C\sum_{k=1}^{N_r}[(R(\xi^r_k) + R(\hat{\xi}^r_k) + R(\xi^i_k) + R(\hat{\xi}^i_k))] - \sum_{k=1}^{N_r}(\eta_i\xi^r_k + \hat{\eta}_k\hat{\xi}^r_k$$
$$+\tau_k\xi^i_k + \hat{\tau}_k\hat{\xi}^i_k) - \sum_{k=1}^{N_r}\alpha_k(\epsilon + \xi^r_k - Re(y_k) + Re(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}})) - \sum_{k=1}^{N_r}\hat{\alpha}_k(\epsilon + \hat{\xi}^r_k + Re(y_k) - Re(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}))$$
$$-\sum_{k=1}^{N_r}\beta_k(\epsilon + \xi^i_k - Im(y_k) + Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}})) - \sum_{k=1}^{N_r}\hat{\beta}_k(\epsilon + \hat{\xi}^i_k + Im(y_k) - Im(< \mathbf{h}_k, \mathbf{w} >_{\mathbb{H}}))$$
$$s.t. \begin{cases} \eta, \hat{\eta}, \tau, \hat{\tau}\alpha, \hat{\alpha}, \beta, \hat{\beta} \geq 0 \\ \xi^r, \hat{\xi}^r, \xi^i, \hat{\xi}^i \geq 0 \end{cases} \tag{32}$$

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

$$\begin{cases} \frac{\partial\Theta}{\partial\mathbf{w}^*} = \frac{1}{2}\mathbf{w} - \frac{1}{2}\sum_{k=1}^{N_r}\alpha_k\mathbf{h}_k + \frac{1}{2}\sum_{k=1}^{N_r}\hat{\alpha}_k\mathbf{h}_k + \frac{i}{2}(\sum_{k=1}^{N_r}\beta_k\mathbf{h}_k - \sum_{k=1}^{N_r}\hat{\beta}_k\mathbf{h}_k) = 0 \\ \Rightarrow \mathbf{w} = \sum_{k=1}^{N_r}(\alpha_k - \hat{\alpha}_k)\mathbf{h}_k - i\sum_{k=1}^{N_r}(\beta_k - \hat{\beta}_k)\mathbf{h}_k \\ \frac{\partial\Theta}{\partial\xi_k^r} = CR'(\xi_k^r) - \eta_k - \alpha_k = 0 \Rightarrow \eta_k = CR'(\xi_k^r) - \alpha_k \\ \frac{\partial\Theta}{\partial\hat{\xi}_k^r} = CR'(\hat{\xi}_k^r) - \hat{\eta}_k - \hat{\alpha}_k = 0 \Rightarrow \hat{\eta}_k = CR'(\hat{\xi}_k^r) - \hat{\alpha}_k \\ \frac{\partial\Theta}{\partial\xi_k^i} = CR'(\xi_k^i) - \tau_k - \beta_k = 0 \Rightarrow \tau_k = CR'(\xi_k^i) - \beta_k \\ \frac{\partial\Theta}{\partial\hat{\xi}_k^i} = CR'(\hat{\xi}_k^i) - \hat{\eta}_k - \hat{\beta}_k = 0 \Rightarrow \hat{\eta}_k = CR'(\hat{\xi}_k^i) - \hat{\beta}_k \end{cases} \tag{33}$$

Based on (33), we have

$$\begin{aligned} &< \mathbf{h}_i, \mathbf{w} >_{\mathbb{H}} = \sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} + i\sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j) < \mathbf{h}_i, \mathbf{h}_j >_{\mathbb{H}} \\ &= \sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\mathbf{h}_i\mathbf{h}_j^H + i\sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\mathbf{h}_i\mathbf{h}_j^H \\ &= \sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{ij}^r - \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\mathbf{K}_{ij}^i + i(\sum_{j=1}^{N_r}(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{ij}^i + \sum_{j=1}^{N_r}(\beta_j - \hat{\beta}_j)\mathbf{K}_{ij}^r), \end{aligned} \tag{34}$$

$$\begin{aligned} ||\mathbf{w}||_{\mathbb{H}}^2 &= \sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\alpha_i - \hat{\alpha}_i)\mathbf{h}_i\mathbf{h}_j^H + \sum_{i,j=1}^{N_r}(\beta_i - \hat{\beta}_i)(\beta_i - \hat{\beta}_i)\mathbf{h}_i\mathbf{h}_j^H \\ &+ i(\sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j)\mathbf{h}_i\mathbf{h}_j^H - \sum_{i,j=1}^{N_r}(\alpha_i - \hat{\alpha}_i)(\beta_j - \hat{\beta}_j)\mathbf{h}_j\mathbf{h}_i^H) \end{aligned} \tag{35}$$

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

$$\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i = -\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ji}^i = -\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i, \tag{36}$$

Therefore

$$\sum_{i,j}^{l} r_i r_j \mathbf{K}_{ij}^i = 0, \tag{37}$$

Based on (37), (35) can be changed to

$$||\mathbf{w}||_{\mathbb{H}}^2=\sum_{i,j=1}^{N_r}(\alpha_i-\hat{\alpha}_i)(\alpha_i-\hat{\alpha}_i)\mathbf{K}_{ij}^r+\sum_{i,j=1}^{N_r}(\beta_i-\hat{\beta}_i)(\beta_i-\hat{\beta}_i)\mathbf{K}_{ij}^r-2\sum_{i,j=1}^{N_r}(\alpha_i-\hat{\alpha}_i)(\beta-\hat{\beta}_i)\mathbf{K}_{ij}^i. \quad (38)$$

Apply (33), (34), (37) and (38) to (32), the final form of Lagrange duality can be obtained

$$\begin{aligned}
&maximize \quad \Theta=-\frac{1}{2}[\sum_{i,j}^{N_r}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\mathbf{K}_{ij}^r+\sum_{i,j}^{N_r}(\beta_i-\hat{\beta}_i)(\beta_j-\hat{\beta}_j)\mathbf{K}_{ij}^r]\\
&-\sum_{i}^{N_r}(\alpha_i+\hat{\alpha}_i+\beta+\hat{\beta}_i)\epsilon+[\sum_{i=1}^{N_r}(\alpha_i-\hat{\alpha}_i)Re(y_i)+\sum_{i=1}^{N_r}(\beta_i-\hat{\beta}_i)Im(y_i)]\\
&+C\sum_{i}^{N_r}(\tilde{R}(\xi_i^r)+\tilde{R}(\hat{\xi}_i^r)+\tilde{R}(\xi_i^i)+\tilde{R}(\hat{\xi}_i^i))\\
&\begin{cases}0\le\alpha(\hat{\alpha})\le C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r))\\ 0\le\beta(\hat{\beta})\le C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i))\\ \xi^r(\hat{\xi}^r)\ge 0\\ \xi^i(\hat{\xi}^i)\ge 0\end{cases}
\end{aligned} \quad (39)$$

which can be divided into 2 independent regression task,

$$\begin{aligned}
&maximize \quad \Theta^r=-\frac{1}{2}\sum_{i,j}^{N_r}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\mathbf{K}_{ij}^r-\sum_{i=1}^{N_r}(\alpha_i+\hat{\alpha}_i)\epsilon+\sum_{i=1}^{N_r}(\alpha_i-\hat{\alpha}_i)Re(y_i)+C\sum_{i=1}^{N_r}(\tilde{R}(\xi_i^r)\\
&+\tilde{R}(\hat{\xi}_i^r))\\
&\begin{cases}0\le\alpha(\hat{\alpha})\le C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r))\\ \xi^r(\hat{\xi}^r)\ge 0\end{cases}
\end{aligned} \quad (40)$$

$$\begin{aligned}
&maximize \quad \Theta^i=-\frac{1}{2}\sum_{i,j}^{N_r}(\beta_i-\hat{\beta}_i)(\beta_j-\hat{\beta}_j)\mathbf{K}_{ij}^r-\sum_{i=1}^{N_r}(\beta_i+\hat{\beta}_i)\epsilon+\sum_{i=1}^{N_r}(\beta_i-\hat{\beta}_i)Im(y_i)+C\sum_{i=1}^{N_r}(\tilde{R}(\xi_i^i)\\
&+\tilde{R}(\hat{\xi}_i^i))\\
&\begin{cases}0\le\beta(\hat{\beta})\le C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i))\\ \xi^i(\hat{\xi}^i)\ge 0\end{cases}
\end{aligned} \quad (41)$$

The alternate form can be written as

$$
\begin{aligned}
&maximize \quad \Theta^r = -\frac{1}{2}(\alpha-\hat{\alpha})^T\mathbf{K}^r(\alpha-\hat{\alpha}) + Re(\mathbf{y})^T(\alpha-\hat{\alpha}) - \epsilon(\mathbf{e}^T(\alpha+\hat{\alpha})) + C(\mathbf{e}^T(\tilde{R}(\xi^r)-\tilde{R}(\hat{\xi}^r)))\\
&\begin{cases} 0 \leq \alpha(\hat{\alpha}) \leq C\tilde{R}(\xi^r)(\tilde{R}(\hat{\xi}^r)) \\ \xi^r(\hat{\xi}^r) \geq 0 \end{cases}
\end{aligned} \tag{42}
$$

$$
\begin{aligned}
&maximize \quad \Theta^i = -\frac{1}{2}(\beta-\hat{\beta})^T\mathbf{K}^r(\beta-\hat{\beta}) + Im(\mathbf{y})^T(\beta-\hat{\beta}) - \epsilon(\mathbf{e}^T(\beta+\hat{\beta})) + C(\mathbf{e}^T(\tilde{R}(\xi^i)-\tilde{R}(\hat{\xi}^i)))\\
&\begin{cases} 0 \leq \beta(\hat{\beta}) \leq C\tilde{R}(\xi^i)(\tilde{R}(\hat{\xi}^i)) \\ \xi^i(\hat{\xi}^i) \geq 0 \end{cases}
\end{aligned} \tag{43}
$$

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 (42) and (43) 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 condition. Therefore, in large MIMO scenario, our CSVR-MIMO detector can save half of the cost in kernel matrix computation.

V. Work Set Selection and Solver

(39) 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 (40) and (41) 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 (42) can be changed to:

$$\begin{aligned}&maximize \quad \Theta^r = -\frac{1}{2}[(\alpha-\hat{\alpha})_S^T \mathbf{K}_{SS}^r (\alpha-\hat{\alpha})_S + 2(\alpha-\hat{\alpha})_N^T \mathbf{K}_{NS}^r (\alpha-\hat{\alpha})_S] + Re(\mathbf{y})_S^T(\alpha-\hat{\alpha})_S - \\ &\epsilon(\mathbf{e}^T(\alpha+\hat{\alpha})_S) - \frac{1}{2}(\alpha-\hat{\alpha})_N^T \mathbf{K}_{NN}^r (\alpha-\hat{\alpha})_N + Re(\mathbf{y})_N^T(\alpha-\hat{\alpha})_N - \epsilon(\mathbf{e}^T(\alpha+\hat{\alpha})_N) \\ &+C(\mathbf{e}^T(\tilde{R}(\xi^r) - \tilde{R}(\hat{\xi}^r))),\end{aligned} \tag{44}$$

Where $\mathbf{K}^r = \begin{bmatrix} \mathbf{K}_{SS}^r & \mathbf{K}_{SN}^r \\ \mathbf{K}_{NS}^r & \mathbf{K}_{NN}^r \end{bmatrix}$, $\mathbf{K}_{SN}^r = \mathbf{K}_{NS}^r$ and $\alpha_S \in \mathbb{R}^{N_r}$ denotes the vector with the components that do not belong to S zeroed. Therefore the sub-optimization problem is formulated as

$$maximize \quad \Theta_S^r = -\frac{1}{2}[(\alpha-\hat{\alpha})_S^T \mathbf{K}_{SS}^r (\alpha-\hat{\alpha})_S] + [Re(\mathbf{y})_S^T - (\alpha-\hat{\alpha})_N^T \mathbf{K}_{NS}^r](\alpha-\hat{\alpha})_S - \epsilon[\mathbf{e}_S^T(\alpha+\hat{\alpha})_S], \tag{45}$$

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

$$maximize \quad \bigtriangledown \Theta_S = \Theta((\alpha + \delta_S \mathbf{e}_S), (\hat{\alpha} + \hat{\delta}_S \mathbf{e}_S)) - \Theta(\alpha, \hat{\alpha}), \tag{46}$$

where $\delta_S = \alpha_S^{new} - \alpha_S$, the gain in (46) can be written as

$$\bigtriangledown\Theta_S^r = -\frac{1}{2}[(\delta-\hat{\delta})_S^T \mathbf{K}_{SS}^r(\delta-\hat{\delta})_S + 2(\alpha-\hat{\alpha})_S^T \mathbf{K}_{SS}^r(\delta-\hat{\delta})_S] + [Re(\mathbf{y})_S^T - (\alpha-\hat{\alpha})_N^T \mathbf{K}_{NS}^r](\delta-\hat{\delta})_S$$
$$-\epsilon(\delta+\hat{\delta})_S = -\frac{1}{2}(\delta-\hat{\delta})_S^T \mathbf{K}_{SS}^r(\delta-\hat{\delta})_S + [Re(\mathbf{y})_S^T - (\alpha-\hat{\alpha})^T \mathbf{K}_S^r](\delta-\hat{\delta})_S - \epsilon(\delta+\hat{\delta})_S \quad (47)$$

In (47), we use

$$(\alpha-\hat{\alpha})_S^T \mathbf{K}_{SS}^r + (\alpha-\hat{\alpha})_N^T \mathbf{K}_{NS}^r = [(\alpha-\hat{\alpha})_S^T, (\alpha-\hat{\alpha})_N^T] \begin{bmatrix} \mathbf{K}_{SS}^r \\ \mathbf{K}_{NS}^r \end{bmatrix} = (\alpha-\hat{\alpha})^T \mathbf{K}_S^r, \quad (48)$$

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 (47) can be rewritten as

$$\bigtriangledown\Theta_S^r = -\frac{1}{2}(\delta-\hat{\delta})_S^T \mathbf{K}_{SS}^r(\delta-\hat{\delta})_S + (\Phi_S^r)^T(\delta-\hat{\delta})_S - \epsilon(\delta+\hat{\delta})_S \quad (49)$$

In our regression model, the offset is omitted, therefore different from SMO type algorithms, there is no linear equation constraint as shown in (20), 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, which can approximate the effect of optimal 2-D work set selection. There is only $O(n)$ searching times required for the former one while the later one need $O(n^2)$ 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 (46), let α_1 denotes the dual variable that is chosen to be updated. The sub optimization problem is formulated as

$$maximize \quad \Theta_1^r = -\frac{1}{2}((\alpha_1 - \hat{\alpha}_1)^{new})^2\mathbf{K}_{11}^r - (\alpha_1 - \hat{\alpha}_1)^{new}\sum_{j=2}^{N_r}\mathbf{K}_{1j}^r(\alpha_j - \hat{\alpha}_j) + Re(y_1)(\alpha_1 - \hat{\alpha}_1)^{new}$$
$$-\epsilon(\alpha_1 + \hat{\alpha}_1)^{new}, \tag{50}$$

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

$$\begin{aligned}
&\frac{\partial\Theta_1^r}{\partial\alpha_1} = -(\alpha_1 - \hat{\alpha}_1)^{new}\mathbf{K}_{11}^r - \sum_{j=2}^{N_r}(\alpha_j - \hat{\alpha}_j)^{new}\mathbf{K}_{1j}^r + Re(y_1) - \epsilon = 0 \\
&\Rightarrow \alpha_1^{new} = \alpha_1 + \frac{\Phi_1^r - \epsilon}{\mathbf{K}_{11}^r},
\end{aligned} \tag{51}$$

$$\begin{aligned}
&\frac{\partial\Theta_1^r}{\partial\hat{\alpha}_1} = (\alpha_1 - \hat{\alpha}_1)^{new}\mathbf{K}_{11}^r + \sum_{j=2}^{N_r}(\alpha_j - \hat{\alpha}_j)^{new}\mathbf{K}_{1j}^r - Re(y_1) - \epsilon = 0 \\
&\Rightarrow \hat{\alpha}_1^{new} = \hat{\alpha}_1 - \frac{\Phi_1^r + \epsilon}{\mathbf{K}_{11}^r}
\end{aligned} \tag{52}$$

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

$$\begin{cases}
(C\tilde{R}(\xi^r) - \alpha)\xi^r = 0 \\
(C\tilde{R}(\hat{\xi}^r) - \hat{\alpha})\hat{\xi}^r = 0 \\
\alpha(\epsilon + \xi^r - Re(y) + \langle \mathbf{h}, \mathbf{w} \rangle_{\mathbb{H}}) = 0 \\
\hat{\alpha}(\epsilon + \hat{\xi}^r + Re(y) - \langle \mathbf{h}, \mathbf{w} \rangle_{\mathbb{H}}) = 0
\end{cases} \tag{53}$$

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

$$\alpha^{new \quad clipped} = [\alpha^{new}]_0^{C\tilde{R}(\xi^r)} \tag{54}$$

$$\hat{\alpha}^{new \quad clipped} = [\hat{\alpha}^{new}]_0^{C\tilde{R}(\hat{\xi}^r)} \tag{55}$$

where $[]_a^b$ denotes to function

$$[x]_a^b = \begin{cases} a & if \quad x \leq a \\ x & if \quad a < x < b \\ b & if \quad x \geq b \end{cases} \tag{56}$$

Based on (49), The gain of objective function respect to ith dual variable pair is

$$\begin{aligned} \bigtriangledown\Theta_i^r &= \Theta^r((\alpha_i + \delta_i \mathbf{e}_i), (\hat{\alpha}_i + \hat{\delta}_i \mathbf{e}_i)) - \Theta^r(\alpha, \hat{\alpha}) \\ &= -\frac{1}{2}(\delta_i - \hat{\delta}_i)^2 \mathbf{K}_{ii}^r + \Phi_1^r(\delta_i - \hat{\delta}_i) - \epsilon(\delta_i + \hat{\delta}_i) \\ &= (\delta_i - \hat{\delta}_i)[-\frac{1}{2}(\delta_i - \hat{\delta}_i)\mathbf{K}_{ii}^r + \Phi_i^r - \epsilon\frac{\delta_i + \hat{\delta}_i}{\delta_i - \hat{\delta}_i}], \end{aligned} \tag{57}$$

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 (50), that is

$$1 = arg_{(i=1,\ldots,N_r)} \max \bigtriangledown\Theta_i^r, \tag{58}$$

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 (45), 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 (45), the sub objective function can be written as

$$\begin{aligned} &maximize \quad \Theta_{1,2}^r = -\frac{1}{2}[(\alpha_1 - \hat{\alpha}_1)^2 \mathbf{K}_{11}^r + (\alpha_2 - \hat{\alpha}_2)^2 \mathbf{K}_{22}^r + 2(\alpha_1 - \hat{\alpha}_1)(\alpha_2 - \hat{\alpha}_2)\mathbf{K}_{12}^r] - \\ &(\alpha_1 - \hat{\alpha}_1)\sum_{j \neq 1,2}^{N_r}(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{1j}^r - (\alpha_2 - \hat{\alpha}_2)\sum_{j \neq 1,2}^{N_r}(\alpha_j - \hat{\alpha}_j)\mathbf{K}_{2j}^r + Re(y_1)(\alpha_1 - \hat{\alpha}_1) + Re(y_2)(\alpha_2 - \hat{\alpha}_2) \\ &-\epsilon(\alpha_1 + \hat{\alpha}_1 + \alpha_2 + \hat{\alpha}_2), \end{aligned} \tag{59}$$

Based on (59), the partial derivative of $\Theta^r_{1,2}$ with respect to α_1 is

$$\frac{\partial\Theta^r_{1,2}}{\partial\alpha_1} = -(\alpha_1-\hat{\alpha}_1)\mathbf{K}^r_{11} - (\alpha_2-\hat{\alpha}_2)\mathbf{K}^r_{12} - \sum_{j\neq 1,2}^{N_r}(\alpha_j-\hat{\alpha}_j)\mathbf{K}^r_{1j} + Re(y_1) - \epsilon =$$
$$-(\alpha_1-\hat{\alpha}_1)\mathbf{K}^r_{11} + Re(y_1) - \sum_{j\neq 1}^{N_r}(\alpha_j-\hat{\alpha}_j)\mathbf{K}^r_{1j} - \epsilon$$
$$\Rightarrow \alpha_1^{new} = \alpha_1 + \frac{(\Phi^r_1-\epsilon)}{\mathbf{K}^r_{11}}, \tag{60}$$

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

$$\hat{\alpha}_1^{new} = \hat{\alpha}_1 - \frac{(\Phi^r_1+\epsilon)}{\mathbf{K}^r_{11}}, \tag{61}$$

$$\alpha_2^{new} = \alpha_2 + \frac{(\Phi^r_2-\epsilon)}{\mathbf{K}^r_{22}}, \tag{62}$$

$$\hat{\alpha}_2^{new} = \hat{\alpha}_2 - \frac{(\Phi^r_2+\epsilon)}{\mathbf{K}^r_{22}}, \tag{63}$$

It is obviously the dual variables in 2-D solver have the same update rule as that of 1-D solver. Based on (49), assume the ith and jth dual variable pair are chosen, the gain of 2-D solver objective function can be written as

$$\bigtriangledown\Theta^r_{ij} = -\frac{1}{2}[(\delta_i-\hat{\delta}_i)^2\mathbf{K}^r_{ii} + (\delta_j-\hat{\delta}_j)^2\mathbf{K}^r_{jj} + 2(\delta_i-\hat{\delta}_i)(\delta_j-\hat{\delta}_j)\mathbf{K}^r_{ij}] + \Phi^r_i(\delta_i-\hat{\delta}_i) + \Phi^r_j(\delta_j-\hat{\delta}_j)$$
$$-\epsilon(\delta_i+\hat{\delta}_i+\delta_j+\hat{\delta}_j), \tag{64}$$

recall the gain of objective function of 1-D solver in (57), we obtain

$$\bigtriangledown\Theta^r_{ij} = \bigtriangledown\Theta^r_i + \bigtriangledown\Theta^r_j - (\delta_i-\hat{\delta}_i)(\delta_j-\hat{\delta}_j)\mathbf{K}^r_{ij},$$

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

Recall (**??**)

$$\phi_{om} = \prod_{i=1}^{N_t}\frac{|r_{ii}|^2}{|r_{ii}|^2 + \sum_{j<i}|r_{ji}|^2}. \tag{65}$$

All the components in $\mathbf{R}$ are independently distributed and $r_{ji} \sim \mathbb{C}N(0,1)$, $|r_{ii}|^2 \sim Gamma(N_r - i+1, 1)$ [15]. Because $|r_{ji}| \sim Rayleigh(1/\sqrt{2})$, $\sum_{j<i} |r_{ji}|^2 \sim Gamma(i-1, 1)$. Defining $\alpha_i = \sum_{j<i} |r_{ji}|^2$ and $\beta_i = |r_{ii}|^2$, α_i and β_i are mutually independent, therefore (**??**) can be rewritten as

$$\phi_{om} = \prod_{i=1}^{N_t} \frac{\beta_i}{\beta_i + \alpha_i}, \tag{66}$$

From [16], if $X \sim Gamma(k_1, \theta)$ and $Y \sim Gamma(k_2, \theta)$, then $\frac{X}{X+Y} \sim B(k_1, k_2)$, where B denotes Beta distribution. Therefore $\frac{\beta_i}{\beta_i+\alpha_i} \sim B(k_1^i, k_2^i)$, where $k_1^i = N_r - i + 1$, $k_2^i = i - 1$. we define $\eta_i = \frac{\beta_i}{\beta_i+\alpha_i}$, it is obvious that η_i are independently distributed. Based on (66), we have

$$\phi_{om} = \prod_{i=1}^{N_t} \eta_i. \tag{67}$$

Therefore the density function of ϕ_{om} can be defined as

$$f_{\phi_{om}}(x) = \frac{1}{x} \sum_{\mathbf{j}} (\prod_{i=1}^{N_t} c(k_1^i k, k_2^i, j^i)) f(-\ln(x)|\mathbf{k_1} + \mathbf{j}), \tag{68}$$

where $\sum_{\mathbf{j}} = \sum_{j^1} \sum_{j^2} \cdots \sum_{j^{N_t}}$, the range of $j^i \in [0, k_2^i - 1]$, $c(k_1^i, k_2^i, j_i) = (-1)^{j^i} \binom{k_2^i - 1}{j^i} [(k_1^i + k_2^i)\mathbb{B}(k_1^i, k_2^i)]^{-1}$, $\mathbb{B}(\alpha, \beta)$ denotes beta function. $f(-\ln(x)|\mathbf{k_1}+\mathbf{j}) = (\prod_{i=1}^{N_t}(k_1^i + j^i)) \sum_{i=1}^{N_t} [exp((k_1^i + j^i)\ln(x)) / \prod_{j=1 j\neq i}^{N_t} (k_1^j + j^j - k_1^i - j^i)]$. $\mathbf{k_1} + \mathbf{j} = [k_1^1 + j^1, \cdots k_1^{N_t - 1} + j^{N_t - 1}, k_1^{N_t} + j^{N_t}]$. Proof: see Appendix C.

Consider logarithmic expectation of ϕ_{om}, we have

$$E[\ln(\phi_{om})] = \sum_{i=1}^{N_t} E[\ln(\eta_i)], \tag{69}$$

where $E[\ln(\eta_i)] = \psi(k_1^i) - \psi(k_1^i + k_2^i)$, thus we have

$$E[\ln(\phi_{om})] = \sum_{i=1}^{N_t} \psi(N_r - i + 1) - N_t \psi(N_r). \tag{70}$$

we can find (70) is consistent with (**??**).

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}, $E[\ln(\phi_{om})]_{em}$, 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 (70) 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

VIII. Hyperparameter Setting

IX. Computer Simulations

X. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Appendix C

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 [15]:

$$f(\mathbf{A}) = \{\tilde{\Gamma}_m(n)det(\mathbf{\Sigma})^n\}^{-1}det(\mathbf{A})^{n-m}etr(-\mathbf{\Sigma}^{-1}\mathbf{A}), \tag{71}$$

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

$$\tilde{\Gamma}_m(\beta) = \pi^{\frac{m(m-1)}{2}} \prod_{i=1}^{m} \Gamma(\beta - i + 1) \quad Re(\beta) > m - 1. \tag{72}$$

Furthermore, from [15], we have

$$\tilde{\Gamma}_m(\beta) = \int_{\mathbf{X}=\mathbf{X}^H>0} etr(-\mathbf{X})det(\mathbf{X})^{\beta-m}d\mathbf{X} \quad Re(\beta) > m - 1. \tag{73}$$

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

$$\begin{aligned} E[\ln(det(\mathbf{A}))] &= \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))f(\mathbf{A})d\mathbf{A} \\ &= \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))\{\tilde{\Gamma}_m(n)det(\mathbf{\Sigma})^n\}^{-1}det(\mathbf{A})^{n-m}etr(-\mathbf{\Sigma}^{-1}\mathbf{A})d\mathbf{A} \\ &= \frac{det(\mathbf{\Sigma})^{-n}}{\tilde{\Gamma}_m(n)} \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))det(\mathbf{A})^{n-m}etr(-\mathbf{\Sigma}^{-1}\mathbf{A})d\mathbf{A}, \end{aligned} \tag{74}$$

if $\mathbf{\Sigma} = \mathbf{I}$, (74) can be written as

$$E[\ln(det(\mathbf{A}))] = \frac{1}{\tilde{\Gamma}_m(n)} \int_{\mathbf{A}=\mathbf{A}^H>0} \ln(det(\mathbf{A}))det(\mathbf{A})^{n-m}etr(-\mathbf{A})d\mathbf{A}. \tag{75}$$

Because $\frac{d}{dn}[det(\mathbf{A})]^{n-m} = \ln(det(\mathbf{A}))det(\mathbf{A})^{n-m}$, (75) can be rewritten as

$$E[\ln(det(\mathbf{A}))] = \frac{1}{\tilde{\Gamma}_m(n)} \frac{d}{dn} \int_{\mathbf{A}=\mathbf{A}^H>0} etr(-\mathbf{A})det(\mathbf{A})^{n-m}d\mathbf{A}, \tag{76}$$

using (73), (76) can be rewritten as

$$E[\ln(\mathbf{A})] = \frac{\tilde{\Gamma}'_m(n)}{\tilde{\Gamma}_m(n)}. \tag{77}$$

Based on (72), we have

$$\tilde{\Gamma}'_m(n) = \pi^{\frac{m(m-1)}{2}} \sum_{i=1}^{m} [\Gamma'(n-i+1) \prod_{j=1, j\neq i}^{m} \Gamma(n-j+1)], \tag{78}$$

Thus we have

$$E[\ln(det(\mathbf{A}))] = \frac{\tilde{\Gamma}'_m(n)}{\tilde{\Gamma}_m(n)} = \sum_{i=1}^{m} \frac{\Gamma'(n-i+1)}{\Gamma(n-i+1)} = \sum_{i=1}^{m} \psi(n-i+1), \tag{79}$$

where ψ denotes Digamma function.

APPENDIX D

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

$$f(x, k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k}. \tag{80}$$

Thus we have

$$E[\ln(x)] = \frac{1}{\Gamma(k)} \int_0^\infty \ln(x) x^{k-1} e^{-x/\theta} \theta^{-k} dx, \tag{81}$$

define $z = x/\theta$ and since $\Gamma(k) = \int_0^\infty x^{k-1} e^{-x} dx$, (81) can be rewritten as

$$E[\ln(x)] = \ln(\theta) + \frac{1}{\Gamma(k)} \int_0^\infty \ln(z) z^{k-1} e^{-z} dz. \tag{82}$$

Because $\frac{d(z^{k-1})}{dk} = \ln(z) z^{k-1}$, (82) 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 E

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

$$f(x_i) = \frac{1}{\mathbb{B}(k_1^i, k_2^i)} x_i^{k_1^i - 1}(1 - x_i)^{k_2^i - 1}, \tag{83}$$

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

$$f_{y_i}(\rho) = |\frac{dy_i}{dx_i}|^{-1} f_{x_i}(g^{-1}(\rho)) = \frac{1}{\mathbb{B}(k_1^i, k_2^i)} e^{-k_1^i \rho}(1 - e^{-\rho})^{k_2^i - 1}. \tag{84}$$

where (84) can be alternatively expressed as [17]

$$f_{y_i}(\rho) = \sum_{j^i=0}^{k_2^i - 1} c(k_1^i, k_2^i, j^i)(k_1^i + j^i) exp(-(k_1^i + j^i)\rho), \tag{85}$$

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 [17], if $a_1, a_2, \cdots a_n$ are independent exponentially distributed random variables, with pdf given by

$$t_i exp(-t_i a_i) \tag{86}$$

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

$$f(a|\mathbf{t}) = \prod_{i=1}^n t_i \sum_{i=1}^n [exp(-t_i a) / \prod_{j=1 j\neq i}^{j=n} (t_j - t_i)], \tag{87}$$

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 (87), the pdf of y is given by

$$f_y(m) = \sum_{\mathbf{j}} \{[\prod_{i=1}^n c(k_1^i, k_2^i, j^i)] f(m|\mathbf{k_1} + \mathbf{j})\}, \tag{88}$$

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

$$f_U(u) = |\frac{du}{dy}|^{-1} f_y(-\ln(u)) = \frac{1}{u} \sum_{\mathbf{j}} \{[\prod_{i=1}^{n} c(k_1^i, k_2^i, j^i)] f(-\ln(u)|\mathbf{k_1} + \mathbf{j})\}. \tag{89}$$

REFERENCES

[1] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *Information Theory, IEEE Transactions on*, vol. 50, no. 9, pp. 1893–1909, 2004.

[2] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[3] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Sbornik: Mathematics*, vol. 1, no. 4, pp. 457–483, 1967.

[4] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Communications and Information theory*, vol. 1, no. 1, pp. 1–182, 2004.

[5] M. Wu, B. Yin, A. Vosoughi, C. Studer, J. R. Cavallaro, and C. Dick, "Approximate matrix inversion for high-throughput data detection in the large-scale MIMO uplink," in *Circuits and Systems (ISCAS), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 2155–2158.

[6] J. C. Fricke, M. Sandell, J. Mietzner, and P. A. Hoeher, "Impact of the Gaussian approximation on the performance of the probabilistic data association MIMO decoder," *EURASIP Journal on Wireless Communications and Networking*, vol. 2005, no. 5, pp. 796–800, 1900.

[7] D. Pham, K. R. Pattipati, P. K. Willett, and J. Luo, "A generalized probabilistic data association detector for multiple antenna systems," *IEEE Communications Letters*, vol. 8, no. 4, pp. 205–207, 2004.

[8] 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.

[9] 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.

[10] T. Lakshmi Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing (CHEMP) receiver in large-scale MIMO systems," 2013.

[11] B. Farhang-Boroujeny, H. Zhu, and Z. Shi, "Markov chain Monte Carlo algorithms for CDMA and MIMO communication systems," *Signal Processing, IEEE Transactions on*, vol. 54, no. 5, pp. 1896–1909, 2006.

[12] T. Datta, N. Ashok Kumar, A. Chockalingam, and B. S. Rajan, "A novel MCMC algorithm for near-optimal detection in large-scale uplink mulituser MIMO systems," in *Information Theory and Applications Workshop (ITA), 2012*. IEEE, 2012, pp. 69–77.

[13] X. Ma and W. Zhang, "Performance analysis for MIMO systems with lattice-reduction aided linear equalization," *Communications, IEEE Transactions on*, vol. 56, no. 2, pp. 309–318, 2008.

[14] A. Papoulis, "Stochastic processes," *McGra. w*, 1996.

[15] D. K. Nagar and A. K. Gupta, "Expectations of functions of complex Wishart matrix," *Acta applicandae mathematicae*, vol. 113, no. 3, pp. 265–288, 2011.

[16] A. K. Gupta and S. Nadarajah, *Handbook of beta distribution and its applications*. CRC Press, 2004.

[17] R. Bhargava and C. Khatri, "The distribution of product of independent beta random variables with application to multivariate analysis," *Annals of the Institute of Statistical Mathematics*, vol. 33, no. 1, pp. 287–296, 1981.