A Low-Complexity Soft-Output wMD Decoding for Uplink MIMO Systems with One-Bit ADCs

Seonho Kim*, Namyoon Lee†, and Song-Nam Hong*

*Ajou University, Suwon, Korea, email: {shkim1005, snhong}@ajou.ac.kr
†POSTECH, Pohang, Korea, email: nylee@postech.ac.kr

Abstract—This paper considers an uplink multiuser multiple-input-multiple-output (MU-MIMO) system with one-bit analog-to-digital converters (ADCs), in which $K$ users with a single transmit antenna communicate with one base station (BS) with $N_r$ receive antennas. In this system, a novel MU-MIMO detection method, named weighted minimum distance (wMD) decoding, was recently proposed, as a practical approximation of maximum likelihood (ML) detector. Despite of its attractive performance, the wMD decoding has two limitations to be used in practice: i) the hard-decision outputs degrade the performance of a following channel code; ii) the computational complexity grows exponentially with the $K$. To address them, we first present a soft-output wMD decoding that efficiently computes soft metrics (i.e., log-likelihood ratios) from one-bit quantized observations. We then reduce the complexity of the soft-output wMD decoding by introducing hierarchical code partitioning. Simulation results demonstrate that the proposed method significantly outperforms the other MIMO detectors with a comparable complexity.

Index Terms—Multiuser MIMO detection, analog-to-digital converter (ADC), one-bit ADC.

I. INTRODUCTION

The use of a very large number of antennas at the base station (BS), referred to as massive multiple-input-multiple-output (MIMO), is one of the promising techniques to cope with the predicted wireless data traffic explosion [1]-[5]. The massive MIMO can improve the system throughput and energy efficiency [5], [6]. In contrast, it can considerably increase the hardware cost and the radio-frequency (RF) circuit consumption [6]. Among all the components in a RF chain, a high-resolution analog-to-digital converter (ADC) is particularly power-hungry as the power consumption of an ADC is scaled exponentially with the number of quantization bits and linearly with the baseband bandwidth [7], [8]. To overcome this challenge, the use of low-resolution ADCs (e.g., 1~3 bits) for massive MIMO systems has received increasing attention over the past years. The one-bit ADC is particularly attractive because of the lower hardware complexity. In this case, the in-phase and quadrature components of the continuous-valued received signals are quantized separately using simple zero-threshold comparators and there is no need for an automatic gain controller [9]. Despite the benefits of using low-resolution ADCs, it gives rise to numerous technical challenges: i) an accurate channel estimation at the receiver (CSIR) is complicated; ii) conventional MIMO detection methods, developed for linear MIMO systems, yield a poor bit error rates (BERs) as the impact of non-linearity of ADCs was not taken into account.

There have been extensive works on the MIMO detection and channel estimation methods for the uplink MIMO systems with one-bit ADCs [13]-[16]. The optimal maximum likelihood (ML) detection was introduced in [10] and low-complexity methods were also presented in [10], [11], [12]. Also, numerous channel estimation methods using one-bit quantized observations were developed as least-square (LS) based method [13], maximum-likelihood (ML) type method [10], zero-forcing (ZF) type method [10], and Bussgang decomposition based method [14]. Recently, a novel MIMO detection method, named weighted minimum distance (wMD) decoding, was presented by viewing the MIMO detection problem as an equivalent coding problem [16]. The equivalent coding problem is to find a codeword of the spatial-domain code $C$ from the one-bit quantized observations obtained from $2N_r$ parallel channels with unequal channel reliabilities (see Fig. 3), where the code $C$ is not designable but is completely determined as a function of a channel matrix $H$. In this problem, the wMD decoding, as an extension of minimum distance (MD) decoding, was presented by exploiting the distinct channel reliabilities appropriately. Furthermore, it was demonstrated that the wMD decoding achieves the optimal ML performance for a perfect CSIR and is more robust to an inaccurate CSIR than ML detector [16].

Despite of its attractive performance, there are two technical challenges so that the wMD decoding will be adopted in practical communication systems. First, the wMD decoding produces the hard-decision outputs as in the other MIMO detection methods in [10], [15], which degrades the performance of a following channel decoder. Also, the computational complexity is not manageable when the number of active users is large. In this paper, we address the above problems, by presenting a soft-output wMD decoding and by reducing its complexity using hierarchical code partitioning. Our contributions are summarized as follows.

- We propose a soft-output wMD decoding for the uplink MU-MIMO systems with one-bit ADCs. The proposed soft-output wMD decoding produces the soft outputs (e.g., log-likelihood ratios (LLRs)) from one-bit quan-
Fig. 1. Uplink MU-MIMO systems in which each receive antenna at a BS is equipped with one-bit ADCs.

The outline of this paper is as follows. In Section II we describe the system model of uplink MIMO system with one-bit ADCs and review the wMD decoding. In Section III we present a soft-output wMD decoding which efficiently computes soft metrics from one-bit quantized observations. In Section IV a low-complexity (soft-output) wMD decoding is described. In Section V, a low-complexity (soft-output) wMD decoding is introduced. Section VI concludes the paper.

Notation: Lower and upper boldface letters represent column vectors and matrices, respectively. For any $k \in \{0, ..., K-1\}$, we let $g(k) = [b_0, b_1, ..., b_{K-1}]^T$ represent the m-ary expansion of $k$ where $k = b_0m^0 + \cdots + b_{K-1}m^{K-1}$ for $b_i \in \{0, ..., m-1\}$. We also let $g^{-1}(\cdot)$ denote its inverse function. For a vector, $g(\cdot)$ is applied element-wise. Likewise, if a scalar function is applied to a vector, it will be performed element-wise. Re(a) and Im(a) represent the real and complex part of a complex vector $a$, respectively.

II. Preliminaries

In this section, we define an uplink multiuser MIMO system with one-bit ADCs and review the wMD decoding proposed in [10].

A. System Model

We consider a single-cell uplink multiuser MIMO system in which $K$ users with a single-antenna communicate with one BS with an array of $N_r > K$ antennas (see Fig. 1). We use the $t$ to indicate a time-index. Let $w_k[t] \in \mathcal{W} = \{0, ..., m-1\}$ represent the user $k$’s message for $k \in \{1, ..., K\}$, each of which contains log $m$ information bits. We also denote $m$-ary constellation set by $\mathcal{S} = \{s_0, ..., s_{m-1}\}$ with power constraint as

$$\frac{1}{m} \sum_{i=0}^{m-1} ||s_i||^2 = \text{SNR}. \quad (1)$$

Then, the transmitted symbol of user $k$ at time $t$, $\tilde{x}_k(w_k[t])$, is obtained by a modulation function $f : \mathcal{W} \rightarrow \mathcal{S}$ as

$$\tilde{x}_k(w_k[t]) = f(w_k[t]) \in \mathcal{S}. \quad (2)$$

When the $K$ users transmit the symbols $\mathbf{x}(\mathbf{w}[t]) = [\tilde{x}_1(w_1[t]), ..., \tilde{x}_K(w_K[t])]^T$, the discrete-time complex-valued baseband received signal vector at the BS, $\tilde{\mathbf{r}}[t] \in \mathbb{C}^{N_r}$, is given by

$$\tilde{\mathbf{r}}[t] = \tilde{\mathbf{H}}\mathbf{x}(\mathbf{w}[t]) + \tilde{\mathbf{z}}[t], \quad (3)$$

where $\tilde{\mathbf{H}} \in \mathbb{C}^{N_r \times K}$ is the channel matrix between the BS and the $K$ users, i.e., the $i$-th row of $\tilde{\mathbf{H}}$ is the channel vector between the $i$-th receive antenna at the BS and the $K$ users. In addition, $\tilde{\mathbf{z}}[t] = [\tilde{z}_1[t], ..., \tilde{z}_{N_r}[t]]^T \in \mathbb{C}^{N_r}$ is the noise vector whose elements are distributed as circularly symmetric complex Gaussian random variables with zero-mean and unit-variance, i.e., $\tilde{z}_i[t] \sim \mathcal{CN}(0, 1)$. We assume a block fading
channel in which the channel matrix $\mathbf{H}$ remains constant during $T$ time slots (e.g., coherence time). A transmission frame containing $T_c$ time slots is composed of two different types of a frame as a pilot transmission frame and a data transmission frame (see Fig. 2). The first $T_i$ time slots are allocated for the pilot transmission frame and the subsequent $T_d$ time slots are allocated for the data transmission frame, i.e., $T_c = T_i + T_d$. During the pilot transmission frame, the $K$ users send the pilot signals that are known at the BS, while during the data transmission frame, the users send the data signals that convey the information to the BS.

In the MIMO system with one-bit ADCs, each receive antenna of the BS is equipped with RF chain followed by two one-bit ADCs that are applied to each real and imaginary part separately. Let $\text{sign}(\cdot) : \mathbb{R} \to \{0, 1\}$ represent the one-bit ADC quantizer function with

$$
\hat{r}[t] = \text{sign}(\tilde{r}[t]) = \begin{cases} 0 & \text{if } \tilde{r}[t] \geq 0 \\ 1 & \text{if } \tilde{r}[t] < 0. 
\end{cases}
$$

Then, the BS receives the quantized output vector as

$$
\mathbf{r}_t[t] = \text{sign}(\text{Re}(\tilde{\mathbf{r}}_t[t])) \quad \text{and} \quad \tilde{\mathbf{r}}_t = \text{sign}(\text{Im}(\tilde{\mathbf{r}}_t[t])).
$$

For the ease of representation, we rewrite the complex input-output relationship in (5) into the equivalent real representation as

$$
\mathbf{r}_t[t] = \text{sign}(\mathbf{H}_\mathbf{x}(\mathbf{w}[t]) + \mathbf{z}[t]),
$$

where

$$
\mathbf{r}_t[t] = [\mathbf{r}_t^\text{T}[t], \mathbf{\hat{r}}_t[t]^\text{T}], \quad \mathbf{x}(\mathbf{w}[t]) = [\text{Re}(\mathbf{x}(\mathbf{w}[t]))^\text{T}, \text{Im}(\mathbf{x}(\mathbf{w}[t]))^\text{T}],
$$

$$
\mathbf{z}[t] = [\text{Re}(\mathbf{z}[t])^\text{T}, \text{Im}(\mathbf{z}[t])^\text{T}],
$$

and

$$
\mathbf{H} = \begin{bmatrix} \text{Re}(\mathbf{H}) - \text{Im}(\mathbf{H}) \\ \text{Im}(\mathbf{H}) \quad \text{Re}(\mathbf{H}) \end{bmatrix} \in \mathbb{R}^{N \times 2K},
$$

and where $N = 2N_r$. This real system representation will be used in the sequel.

**B. wMD Decoding**

We review the wMD decoding presented in [16]. This method was developed by showing the equivalence of the original MIMO detection problem and a non-linear coding problem (see Fig. 3). The equivalent coding problem consists of the three parts as described below. Since this method is applied symbol-by-symbol, we in this section drop the time-index $t$ for the ease of exposition. It is assumed that, during the channel training phase, a channel matrix $\mathbf{H}$ is estimated at the BS. Then, we will explain the wMD decoding which is performed to decode the users’ messages during the data transmission phase.

**i) Auto-encoding function:** For a given channel matrix $\mathbf{H}$, a code $C$ over a spatial domain is defined as

$$
C = \{c_0, \ldots, c_{mK-1}\},
$$

where each codeword $c_\ell$ is defined as

$$
c_\ell = [\text{sign}(h_1^\text{T}(g(\ell))), \ldots, \text{sign}(h_N^\text{T}(g(\ell)))]^\text{T}.
$$

The code $C$ is a non-linear binary code of length $N$ and code rate $\frac{K}{NK}$. Since this code is completely described as a function of channel matrix $\mathbf{H} = [h_1, \ldots, h_N]^\text{T}$, this code is referred to as a spatial-domain code. Also, we call a channel code time-domain code.

In Fig. 3 the input $\mathbf{q}[t]$ of an effective channel is generated by an auto-encoding function $\mathcal{E} : \{0, \ldots, m - 1\}^K \to C$ as

$$
\mathbf{q} = \mathcal{E}(\mathbf{w}) = c_\ell
$$

where $\ell = g^{-1}(\mathbf{w}) \in \{0, 1, \ldots, mK - 1\}$.

**Example 1:** Consider a $2 \times 2$ MIMO system with one-bit ADC, and each user is assumed to use 4-QAM, i.e., $N_r = 2$, $K = 2$, and $m = 4$. Then, for a given channel matrix $\mathbf{H} \in \mathbb{R}^{4\times 4}$, one can create a code $C = \{c_1, c_2, \ldots, c_{16}\}$ in which the $\ell$-th codeword is defined as

$$
c_\ell = [\text{sign}(h_1^\text{T}(g(\ell))), \ldots, \text{sign}(h_N^\text{T}(g(\ell)))]^\text{T} \in \{0, 1\}^4.
$$

**ii) Effective channel:** As shown in Fig. 3 the effective channel consists of $N$ parallel binary input/output channels with input $\mathbf{q} = [q_1, \ldots, q_N]^\text{T}$ and output $\mathbf{r} = [r_1, \ldots, r_N]^\text{T}$. For the $i$-th subchannel, the transition probabilities, depending on users’ messages $\mathbf{w} = g(\ell)$, are defined as

$$
p_{\ell,i,j} \triangleq P(r_i = j|q_i = c_{\ell,i}),
$$

for $j \in \{0, 1\}$. This is simply computed using Q-function as

$$
p_{\ell,i,j} = \begin{cases} 
\epsilon_{\ell,i} & \text{if } i \neq j \\
1 - \epsilon_{\ell,i} & \text{if } i = j.
\end{cases}
$$

![Fig. 2. Frame structure consisting of the channel training and data transmissions, during a coherence time.](image-url)

![Fig. 3. Description of an equivalent coding problem. Note that an auto-encoding function $\mathcal{E}$ is determined as a function of $\mathbf{H}$ and a one-bit quantization function. Also, the transition probabilities of an effective channel depend on the message vector $\mathbf{w}$ (i.e., asymmetric channel).](image-url)
where $\epsilon_{i,j}$ $\triangleq Q(\|h_i^T x(g(t))\| < 0)$ denotes a cross-probability of the channel $i$ and 

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left( -\frac{x^2}{2} \right) dt.$$ 

**iii) Decoding function:** The wMD decoding was presented in [16] as an extension of a minimum distance (MD) decoding. 

**Definition 1:** A weighted Hamming distance is defined as 

$$d_{wh}(x, y; \alpha) = \sum_{i=1}^{N} 2^{\alpha_i} \mathbb{1}_{\{x_i \neq y_i\}}.$$

where $\alpha = (\alpha_1, ..., \alpha_N)$ denotes a weight vector, $\mathbb{1}_A$ represents an indicator function with $\mathbb{1}_A = 1$ if $A$ is true, and $\mathbb{1}_A = 0$, otherwise. Note that the Hamming distance is a special case of the weighted Hamming distance with equal weights (i.e., $\alpha_i = 1$ for all $i$).

Using the definition, the wMD decoding is performed as 

$$\hat{\ell} = \arg\min_{\ell \in [1:m^K]} d_{wh}(r, c_\ell; \alpha_\ell), \quad (11)$$

where the weights are defined using the channel reliabilities as 

$$\alpha_{i,j} = -\log \left( Q(\|h_i^T x(g(t))\| < 0) \right), \quad (12)$$

for $i \in \{0, 1, ..., m^K - 1\}$. The key idea of the wMD decoding is to allocate a higher belief to the information conveyed from a more reliable channel while MD decoding assigns an identical belief. Also, it was demonstrated in [16] that the wMD decoding outperforms MD decoding due to the use of the weights.

**III. SOFT-OUTPUT WMD DECODING**

Likewise ML and near ML detectors in [10], and supervised-learning based detector in [15], the wMD decoding produces the hard-decision outputs. Inevitably, a hard channel decoder (e.g., bit-flipping decoder) should be employed as in [10]. This approach can yield a non-trivial performance loss compared to using soft channel decoder (e.g., belief-propagation decoder). To overcome this problem, we propose a soft-output wMD decoding which generates a soft metric (e.g., LLR) from one-bit quantized (hard-decision) observation.

We first define the subcode of the $C$ as follows: 

**Definition 2:** Recall that a spatial-domain code $C$ is defined as 

$$C \triangleq \{ c \in \mathcal{C}(w) : w \in W^K \}. \quad (13)$$

For any given user’s message $\{w_k = j\}$ with $j \in W$, the subcode of the $C$ is defined as 

$$C_{\{w_k = j\}} \triangleq \{ c \in \mathcal{C}(w) : w \in W^K, w_k = j \}.$$

Using the above definition, we will compute the a posteriori probabilities (APPs) from the one-bit quantized observation $r[t] = (r_1[t], ..., r_N[t])$, where the APPs are defined as 

$$\mathbb{P}\{w_k[t] = j | r[t]\} : j \in W, k \in \{1, ..., K\}\}.$$ 

We let $w_k[t] = (w_1[t], ..., w_{k-1}[t], w_{k+1}[t], ..., w_K[t])^T$. Then, the APP of the user $k$’s message is computed as 

$$\mathbb{P}(w_k[t] = j | r[t]) = \frac{1}{Z} \sum_{u \in W^{K-1}} \mathbb{P}(w_k[t] = j, w_{k}[t] = u | r[t])$$

\[\begin{align*}
&= \frac{1}{Z} \sum_{u \in W^{K-1}} \mathbb{P}(r[t] | w_{k}[t] = j, w_{k}[t] = u) \\
&= \frac{1}{Z} \sum_{c \in \mathcal{C}[w_{k}[t] = j]} \mathbb{P}(r[t] | c[t]),
\end{align*}\]

for $j \in W$, where (a) is from the Bayes’ rule, (b) is from Definition 2, $\mathbb{P}(r[t] | c[t])$ is defined in [9], and $Z$ denotes a normalization factor such that 

$$\sum_{j \in W} \mathbb{P}(w_k[t] = j | r[t]) = 1. \quad (16)$$

Using the weighted Hamming distance in Definition 1, the [15] can be approximately computed as 

$$\mathbb{P}(w_k[t] = j | r[t]) \approx \frac{1}{Z} \exp \left( -\sum_{c \in \mathcal{C}[w_{k}[t] = j]} d_{wh}(r[t], c[t], \alpha_c) \right). \quad (17)$$

Note that the above approximation is very accurate when the crossover probability of each subchannel is smaller than 0.3 [16]. Also, using the well-known approximation as 

$$\exp(x_1 + x_2 + \cdots + x_i) \approx \exp(\max\{x_1, x_2, ..., x_i\}), \quad (18)$$

the (17) can be further simplified as 

$$\mathbb{P}(w_k[t] = j | r[t]) \approx \frac{1}{Z} \exp \left( -\min_{c \in \mathcal{C}[w_{k}[t] = j]} d_{wh}(r[t], c[t], \alpha_c) \right). \quad (19)$$

From the APPs derived in (19) (or (14)), we then compute the soft inputs (e.g., log-likelihood ratios (LLRs)) of a channel decoder. To make an explanation clear, we only consider a 2q-QAM constellation (e.g., $W = \{0, 1, ..., 2^q\}$) for some positive $q$. However, the extension to an arbitrary m-ary constellation is straightforward. Fig. [2] describes the coded system for $q = 2$ (i.e., 4-QAM). Let $[\tau_k[1], ..., \tau_k[n]]$ denote the coded output of the user $k$’s channel encoder. For the ease of notation, we define: 

$$[b]_q \triangleq \sum_{i=1}^{q} b_i 2^{-q-i}, \quad (20)$$

where $b = (b_1, ..., b_q)$ with $b_i \in \{0, 1\}$ for $i = 1, ..., q$. Then, the user $k$’s channel input message at time slot $t$ is obtained as 

$$w_k[t] = ([\tau_k[qt], \tau_k[qt - 1], ..., \tau_k[qt - q + 1]])_q,$$ 

for $t = 1, ..., n/q$, where it is assumed that $n$ is a multiple of $q$. Each user $k$ transmits the $\{w_k[t] : t = 1, ..., n/q\}$ to the BS over the $n/q$ time slots. From the observations $\{r[t] : t = 1, ..., n/2\}$ and using (19), the BS first computes the APPs as 

$$\mathbb{P}(w_k[t] = j | r[t]) : j \in W, t = 1, ..., n/q. \quad (22)$$
Then, it computes the soft inputs (e.g., LLRs) of the channel decoder as

\[ L_{qt}^k\{r[t]\} = \log \frac{\Pr\{r[t] = 1\} - \Pr\{r[t] = 0\}}{\Pr\{r[t] = 0\} - \Pr\{r[t] = 1\}} \]

\[ = \log \frac{\sum b \in \{0,1\} \cdot b \cdot \Pr(w_k[t] = |b|_q)}{\sum b \in \{0,1\} \cdot b \cdot \Pr(w_k[t] = |b|_q)}, \]

for \( i = 1, ..., q \) and \( t = 1, ..., n/q \). This can be simply computed from (18) and (19) as

\[ L_{qt}^{k-1}\{r[t]\} = \min_{c_i \in \cup b \in \{0,1\} \cdot b \cdot C_{|b|_q} = |b|_q} d_w(h\{r[t]\}, c_i, \alpha_k), \]

\[ \quad - \min_{c_i \in \cup b \in \{0,1\} \cdot b \cdot C_{|b|_q} = |b|_q} d_w(h\{r[t]\}, c_i, \alpha_k), \]

for \( i = 1, ..., q \) and \( t = 1, ..., n/q \).

Example 2: When 4-QAM is used, the LLRs are computed from (23) as

\[ L_{2t-1}^k\{r[t]\} = \min_{c_i \in \cup c_{|b|_2} \cdot b \cdot \cup c_{|b|_2} \cdot b} d_w(h\{r[t]\}, c_i, \alpha_k), \]

\[ \quad - \min_{c_i \in \cup c_{|b|_2} \cdot b \cdot \cup c_{|b|_2} \cdot b} d_w(h\{r[t]\}, c_i, \alpha_k), \]

\[ L_{2t}^k\{r[t]\} = \min_{c_i \in \cup c_{|b|_2} \cdot b \cdot \cup c_{|b|_2} \cdot b} d_w(h\{r[t]\}, c_i, \alpha_k), \]

\[ \quad - \min_{c_i \in \cup c_{|b|_2} \cdot b \cdot \cup c_{|b|_2} \cdot b} d_w(h\{r[t]\}, c_i, \alpha_k), \]

for \( t = 1, ..., n/2 \).

IV. A LOW-COMPLEXITY SOFT-OUTPUT WMD DECODING USING HIERARCHICAL CODE PARTITIONING

We observe that the computational complexity of the soft-output WMD decoding as well as WMD decoding is problematic for a large \( K \) as the size of the code \( C \) (i.e., the search-space) grows exponentially with the \( K \). In this section, we present a low-complexity soft-output WMD decoding by introducing hierarchical code structure. Note that the proposed method is directly applied to the WMD decoding. The key idea of the proposed method is that the code \( C \) is partitioned in a hierarchical manner: the code \( C \) is partitioned into the level-1 subcodes and each level-1 subcode is further partitioned into the level-2 subcodes, and so on (see Section IV-A).

This process is referred to as **hierarchical code partitioning**. Leveraging this structure, we can efficiently identify some codewords of the \( C \) that lie inside the sphere centered at the current observation \( r[t] \) with a certain radius, where the reduced code is denoted by \( C_i\{r[t]\} \). This method is reminiscent of a sphere decoding (22), (23) that is developed for conventional MIMO systems. In this sense, the proposed method can be regarded as a sphere decoding for the MIMO systems with one-bit ADCs.

To be specific, the proposed method consists of three parts: i) hierarchical code partitioning; ii) pre-processing; iii) soft-output WMD decoding. During a coherence time, the part i) is performed at once in channel training phase while the parts ii) and iii) are performed at each time slot in data transmission phase (see Fig. 2). The detailed procedures are described as follows.

A. Channel Training Phase

In this phase, the BS first estimates a channel matrix \( \mathbf{H} \) using the \( T_i \) pilot signals where numerous channel estimation methods can be used (see (10) and (14) for details). Using the \( \mathbf{H} \), the BS creates the spatial-domain code \( \mathbf{C} = \{ \mathbf{c}_0, ..., \mathbf{c}_m \} \), defined in (7), and computes the weights (channel reliabilities) of \( N \) parallel channels \( \alpha_k = (\alpha_{k,1}, ..., \alpha_{k,N}) \). Then, the (soft-output) WMD decoding can be performed.

The following procedures are required to perform the low-complexity (soft-output) WMD decoding. For a fixed hierarchical level \( L \geq 1 \), the code \( C \) is partitioned into several subcodes in a hierarchical manner:

- At the level-1, using a vector quantization method, the \( C \) is partitioned into the \( k_1 \) subcodes \( C(1), ..., C(k_1) \) with \( \cup C(1) = C \) and the level-1 subcode is a \( k_1 \)-dimensional vector.
- At the level-2, using a vector quantization method, the \( C(i) \) is partitioned into the \( k_2 \) subcodes \( C(i)(1), ..., C(i)(k_2) \) with \( \cup C(i)(1) = C(i) \) and the level-2 subcode is a \( k_2 \)-dimensional vector.

Fig. 4. The proposed coded architecture for uplink MU-MIMO systems with one-bit ADCs.
Then the (soft-output) wMD decoding is performed using the pre-processing, some unnecessary codewords (having a Hamming distance less than the target distance) are eliminated from the search-space.

- At the level-2, each level-1 subcode \( C_{(i_1)} \) is further partitioned into the \( k_2 \) subcodes \( C_{(i_1,i)} \) for \( i = 1, ..., k_2 \) using the \( k \)-means clustering algorithm. They satisfy the partitioning

\[
\bigcup_{i=1}^{k_2} C_{(i_1,i)} = C_{(i_1)}.
\]

Also, the \( k_2 \) centroids \( \{\mu_{(i_1,i)} : i = 1, 2, ..., k_2\} \) are generated and for each centroid \( \mu_{(i_1,i)} \), the weight vector \( \beta_{(i_1,i)} \) is computed using \( d_{ij} \).

- Generally at the level-\( \ell \), each level-(\( \ell - 1 \)) subcode \( C_{(i_1,i_2,\ldots,i_{\ell-1})} \) is further partitioned into the \( k_\ell \) subcodes \( C_{(i_1,i_2,\ldots,i_{\ell-1},i)} \) for \( i = 1, ..., k_\ell \) and the corresponding \( k_\ell \) centroids \( \{\mu_{(i_1,i_2,\ldots,i_{\ell-1},i)} : i = 1, 2, ..., k_\ell\} \) are generated. Also, for each centroid \( \mu_{(i_1,i_2,\ldots,i_{\ell-1},i)} \), the weight vector \( \beta_{(i_1,i_2,\ldots,i_{\ell-1},i)} \) is computed using \( d_{ij} \).

- Repeatedly perform the above process for \( \ell = 1, ..., L \).

The above process is referred to as hierarchical code partitioning because this process partitions the code \( C \) into the subcodes with the hierarchical structure (see Fig. 5). Note that the resulting subcodes are used during the coherence time (e.g., \( T \) time slots), as shown in Fig. 2.

### B. Data Transmission Phase

In the data transmission, the decoding consists of the two parts as pre-processing and (soft-output) wMD decoding. In the pre-processing, some unnecessary codewords (having a lower probability to be a valid codeword) are precluded, and then the (soft-output) wMD decoding is performed using the reduced code.

1) Pre-processing: As shown in Fig. 6, this process is performed as follows.

- With the weight vector \( \beta_{(i)} \), the weighted Hamming distances between the \( r[t] \) and the level-1 centroids \( \mu_{(1)}, \ldots, \mu_{(k_1)} \) is are computed as

\[
d_i = d_{wh}(\mu_{(i)}, r[t], \beta_{(i)}),
\]

for \( i = 1, \ldots, k_1 \). Sort the \( d_i \)'s in an increasing order and then define the index set containing the first \( q_1 \) indices as \( I_1 = \{i_1, i_2, \ldots, i_{q_1}\} \). In this process, the codewords outside the chosen subcodes are eliminated from the search-space.

- Similarly, with the weight vectors \( \{\beta_{(i_1,i)} : i_1 \in I_1, i = 1, 2, \ldots, k_2\} \), the weighted Hamming distances between the \( r[t] \) and the level-2 centroids \( \{\mu_{(i_1,i)} : i_1 \in I_1, i = 1, 2, \ldots, k_2\} \) are computed, and then the corresponding index set \( I_2 = \{i_1, i_2\} \) is defined. Note that this process further reduces the search-space by ruling out the unnecessary codewords.

- In general, the weighted Hamming distances between the \( r[t] \) and the level-\( \ell \) centroids \( \{\mu_{(i_1,i_2,\ldots,i_{\ell-1},i)} : (i_1, i_2, \ldots, i_{\ell-1}) \in I_{\ell-1}, i = 1, 2, \ldots, k_\ell\} \) are computed with the weight vectors \( \{\beta_{(i_1,i_2,\ldots,i_{\ell-1},i)} : (i_1, i_2, \ldots, i_{\ell-1}) \in I_{\ell-1}, i = 1, 2, \ldots, k_\ell\} \), and the corresponding index set \( I_{\ell} \) is defined.

- Repeatedly perform the above process for \( \ell = 1, 2, \ldots, L \). From the pre-processing, the reduced code \( C_t(r[t]) \subset C \) is obtained as

\[
C_t(r[t]) = \bigcup_{(i_1,i_2,\ldots,i_{L}) \in I_L} C_{(i_1,i_2,\ldots,i_{L})}.
\]

Note that the \( C_t(r[t]) \) depends on the current observation \( r[t] \) and only contains the codewords which are close to the \( r[t] \) in some sense. It is noticeable that in the proposed method, the \( q_{\ell} > 1 \) subcodes can be chosen concurrently for each level \( \ell \). This is to improve the probability that a valid codeword belongs to the \( C_t(r[t]) \), with the expense of the complexity. Therefore, the parameters \( \{\{k_1, \ldots, k_L\}, \{q_1, \ldots, q_L\}\} \) should be carefully chosen by taking the performance-complexity trade-off into account. Also, since he number of chosen subcodes at the level \( \ell \) should be smaller than the remaining subcodes at the level \( \ell - 1 \), the parameters should satisfy the condition of

\[
q_\ell \leq q_{\ell-1} - k_\ell
\]

for \( \ell = 1, \ldots, L \), where \( q_0 = 1 \).
and the star node denotes the received observation. Also, the dashed circles denote the level-1 subcodes and the solid squares denote the level-2 subcodes. After the pre-processing, the black-colored triangle nodes are only remained in the search-space.

Fig. 6. Illustration of the pre-processing when 2-level hierarchical code partitioning is used. The triangle nodes denote the codewords of the C and the star node denotes the received observation. Also, the dashed circles denote the level-1 subcodes and the solid squares denote the level-2 subcodes. After the pre-processing, the black-colored triangle nodes are only remained in the search-space.

Fig. 7. K = 6 and Nr = 64. BER performances of the proposed method according to the choices of \{\{k_1, q_1\}\} when 1-hierarchical level is considered.

by the filled squares) are chosen. After this process, the wMD decoding is performed with the codewords belong to the \(C_{(7,2)} \cup C_{(8,1)}\).

C. Discussion on Computational Complexity

In this section, we discuss the complexity of the low-complexity wMD decoding for each coherence time \(T_c = T_i + T_d\), where the complexity is measured as the number of distance comparisons. Let \(N_{\text{cp}}, N_{\text{pre}}, \text{and } N_{wMD}\) denote the number of distance comparisons required for hierarchical code partitioning, pre-processing, and wMD decoding, respectively. Then, the overall complexity during the coherence time \(T_c\) is given by \(N_{\text{cp}} + T_d(N_{\text{pre}} + N_{wMD})\). Accordingly, the average complexity per time slot is given by

\[
N_{\text{total}} = \frac{1}{T_c} N_{\text{cp}} + \frac{T_d}{T_c} (N_{\text{pre}} + N_{wMD}) \tag{29}
\]

where the above approximation is generally accurate since \(T_d \gg T_i\) and \(T_c \gg 1\). Thus, we assume the \(N_{\text{total}} = N_{\text{pre}} + N_{wMD}\) as the average complexity per time slot.

First, the pre-processing complexity is computed as

\[
N_{\text{pre}} = \sum_{\ell=1}^{L} q_{\ell-1} k_\ell, \tag{31}
\]

where \(q_0 = 1\), since there are the \(q_{\ell-1} k_\ell\) number of centroids for each level \(\ell\). After the pre-processing, the number of the remaining codewords in the search-space is

\[
N_{wMD} = \sum_{(i_1,\ldots,i_{\ell}) \in \mathcal{I}_\ell} |C_{(i_1,\ldots,i_{\ell})}|. \tag{32}
\]

In fact, the \(N_{wMD}\) is not a constant but is determined as a function of a channel matrix \(H\) and an observation \(r[t]\). This is because the \(k\)-means clustering algorithm does not ensure the equi-partitioning of the code \[17\]. Via numerical results, we verified that the average value of \(N_{wMD}\), where the average is performed over a random channel matrix, is very well approximated to the complexity obtained with the assumption of the uniform partitioning as

\[
N_{wMD} \approx |\mathcal{W}|^K \times \frac{q_1}{k_1} \times \frac{q_2}{q_1 k_2} \cdots \times \frac{q_L}{q_{L-1} k_L} = |\mathcal{W}|^K \frac{q_L}{\prod_{\ell=1}^{L} k_\ell}. \tag{33}
\]

With this approximation, the average decoding complexity per time slot is given by

\[
N_{\text{total}} = N_{\text{pre}} + N_{wMD} \approx \sum_{\ell=1}^{L} q_{\ell-1} k_\ell + |\mathcal{W}|^K \frac{q_L}{\prod_{\ell=1}^{L} k_\ell}, \tag{34}
\]

which is assumed as the average complexity in the sequel.

Example 4: Consider the uplink MIMO systems with \(K = 8\) and \(N_r = 64\) where 4-QAM is assumed. The overall complexity of wMD decoding is very expensive as \(N_{\text{total}} \approx 65536\). Using the 1-hierarchical level \{\{(32, 8)\}\}, the complexity can be reduced to the 25% of the original complexity as \(N_{\text{total}} \approx 16416\). Also, using the 3-hierarchical level \{\{(32, 4, 4), (8, 8, 8)\}\}, the complexity can be further reduced to the 1.7% of the original complexity as \(N_{\text{total}} \approx 1120\). In Fig. 9 it is shown that the performance obtained with the 3-hierarchical level approaches the optimal performance of the wMD decoding.

V. NUMERICAL RESULTS

We evaluate the performances of the low-complexity (soft-output) wMD decoding. A Rayleigh fading channel is assumed in which each element of a channel matrix \(H\) is drawn from an independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian random variable with zero mean and unit variance. Also, 4-QAM and ZF-type channel estimation method in \[10\] are assumed.

Fig. 8 shows the BER performances of the low-complexity wMD decoding according to the choices of \{\{(k_1, q_1)\}\}, where the parameters are chosen such that the size of the reduced code (i.e., search-space) is equal to 1024. We observe that the BER performance is enhanced by partitioning the code \(C\) into

\[
N_{\text{pre}} + N_{wMD}\]
The coherence time, which is set to be $T$, is relatively small compared to the complexity of the wMD decoding. Also, to reduce the decoding complexity of the (soft-output) wMD decoding, the 2-hierarchical level is considered.

Fig. 9 shows the uncoded BER performances of the proposed method according to the number of hierarchical levels. The subcodes with a smaller size, and the performance gain is unbounded due to the error-floor. From this observation, the best strategy for selecting the $\{(k_1, q_1)\}$ is to choose a larger $k_1$ as long as the complexity of the pre-processing is relatively small compared to the complexity of the wMD decoding.

Fig. 8 shows the BER performances of the proposed method according to the choices of $\{(k_1, q_1)\}$ when 1-hierarchical level is considered.

Fig. 10 shows the BER performances of the various MIMO detection methods where the coded system is formed by concatenating a MIMO detector with a low-density-parity-check (LDPC) code. We adopt a rate 1/2 LDPC code of the blocklength 672 from the IEEE802.11ad standardization [19]. As an LDPC code decoder, the bit-flipping decoder [20] is used for the wMD decoding and ZF-type detector, and the belief-propagation decoder [21] is used for the soft-output wMD decoding. Also, to reduce the complexity of the (soft-output) wMD decoding, the 2-hierarchical level $\{(32, 4), (8, 8)\}$ is assumed. In this example, a block fading duration (i.e., coherence time) is set to be $T_c = T_t + T_d = 1369$ with $T_t = 25$ and $T_d = 1274$, where the two coded outputs of the LDPC code are transmitted during the coherence time. This example shows that the soft-output wMD decoding has a non-trivial performance gain over the wMD decoding (or ML detector) and ZF-type detector with a comparable complexity.

VI. CONCLUSION

We proposed the soft-output wMD decoding which efficiently computes the soft outputs (e.g., log-likelihood ratios) from one-bit quantized observations. This enables to employ soft channel decoder (e.g., belief-propagation decoder) for the MIMO systems with one-bit ADCs. Furthermore, we noticeable that the ML detection with imperfect CSIR severely suffers from the BER degradation especially in the high-SNR regimes due to the impact of the inaccurate CSIR. In contrast, the wMD decoding with imperfect CSIR outperforms the existing MIMO detection techniques and the performance gaps increase as SNR grows. Namely, the wMD decoding is more robust to imperfect CSIR than ML detection although both methods achieve the same optimal performance with perfect CSIR. We notice that the use of 2-hierarchical level can reduce the decoding complexity of the 10% of the original complexity with a small performance loss. Thus, in an imperfect CSIR, the low-complexity wMD decoding can provide a satisfactory performance with a manageable decoding complexity.

In Fig. 11 we compare the low-complexity wMD decoding with the existing MIMO detection methods. A block fading duration (i.e., coherence time) is set to be $T_c = T_t + T_d = 1000$ time slots and the training overhead is set to the 2.5% of the coherence time (i.e., $T_t = 25$). For the comparisons, we consider the ML and ZF detection methods in [10]. It is
presented the low-complexity soft-output wMD decoding by introducing hierarchical code partitioning, which can be regarded as a sphere decoding for the MIMO systems with one-bit ADCs. Finally we demonstrated that the proposed method significantly outperforms the other MIMO detectors with hard-decision outputs, with a comparable complexity. One possible extension is to study the soft-output wMD decoding for a slowly varying channel, in which we may reduce the channel training overhead by updating the spatial-domain code and the weights from the previous ones, rather than newly constructing them.

REFERENCES

[1] T. L. Marzetta, “Noncooperative cellular wireless with unlimited numbers of base station antennas,” *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590-3600, Nov. 2010.

[2] A. Adhikary, J. Nam, J.-Y. Ahn and G. Caire, “Joint Spatial Division and Multiplexing - The Large-Scale Array Regime,” *IEEE Trans. Inf. Theory*, vol. 59, pp. 6441-6463, Jun. 2013.

[3] A. Adhikary, E. A. Safadi, M. K. Samimi, R. Wang, G. Caire, T. S. Rappaport and A. F. Molisch, “Joint Spatial Division and Multiplexing for mm-Wave Channels,” *IEEE J. Sel. Commun.*, vol. 32, pp. 1239-1255, May 2014.

[4] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186-195, Feb. 2014.

[5] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, “An overview of massive MIMO: benefits and challenges,” *IEEE J. Sel. Topics Sig. Proc.*, vol. 8, no. 5, pp. 742-758, Oct. 2014.

[6] H. Yang and T. L. Marzetta, “Total energy efficiency of cellular large scale antenna system multiple access mobile networks,” in *Proc. IEEE Online Conf. Green Commun.*, Piscatway, NJ, pp. 27-32, Oct. 2013.

[7] B. Murmann, “ADC Performance Survey 1997-2015,” [Online]. Available: http://web.stanford.edu/ murmann/adcsurvey.html.

[8] A. Mezghani and J. A. Nossek, “Modeling and minimization of transceiver power consumption in wireless networks,” in *Proc. IEEE/ITG WSA*, pp. 1-8, Feb. 2011.

[9] S. Hoyos, B. M. Sadler and G. R. Arce, “Monobit digital receivers for ultrawideband communications,” *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1337-1344, Jul. 2005.

[10] J. Choi, J. Mo and R. W. Heath Jr., “Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs,” *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 2005-2018, May 2016.

[11] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, Jr., “One-bit ADCs in wideband massive MIMO systems with OFDM transmission,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Mar. 2016.

[12] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, Jr., “Uplink performance of wideband massive MIMO with one-bit ADCs,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 2156-2168, Jan. 2017.

[13] C. Risi, D. Persson and E. G. Larsson, “Channel estimation and performance analysis of one-bit massive MIMO systems,” [Online]. Available: http://arxiv.org/abs/1409.07427 Sep. 2016.

[14] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, “Channel estimation and performance analysis of one-bit massive MIMO systems,” [Online]. Available: http://arxiv.org/abs/1609.07427 Sep. 2016.

[15] Y. Jeon, S.-N. Hong, and N. Lee, “Supervised-learning-aided communication framework for massive MIMO systems with low-resolution ADCs,” submitted to *IEEE Trans. Sig. Proc.*, Mar. 2017.

[16] S.-N. Hong, S. Kim, and N. Lee, “A weighted minimum distance decoding for uplink multiuser MIMO systems with low-resolution ADCs,” submitted to *IEEE Trans. Commun.*, Jun. 2017.

[17] S. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, pp. 129-137, Jan. 2003.

[18] C. Leroux, A. J. Raymond, G. Sarkis, and W. J. Gross, “A semi-parallel successive-cancellation decoder for polar codes,” *IEEE Trans. Signal Process.*, vol. 61, pp. 2897-2929, Jan. 2013.

[19] *IEEE Approved Draft Standard for LAN - Specific Requirements - Part II: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - Amendment 3: Enhancements for Very High Throughput in the 60GHz Band*, IEEE P802.11ad/D9.0 Std., Jul. 2012.

[20] K. D. Rao, *Channel Coding Techniques for Wireless Communications*, Springer, 2015.

[21] T. Richardson and R. Urbanke, *Modern coding theory*, Cambridge university press, 2008.

[22] C. P. Schnorr and M. Euchner, “Lattice basis reduction: improved practical algorithms and solving subset sum problems,” *Math. Programming*, vol. 66, pp. 181-191, Sept. 1995.

[23] U. Fincke and M. Pohst, “Improved methods for calculating vectors of short length in a lattice, including a complexity analysis,” *Mathematics of Computation*, vol. 44, pp. 463-471, Apr. 1985.