Tradeoff Between Delay and High SNR Capacity in Quantized MIMO Systems

Abbas Khalili
Department of Electrical and Computer Engineering
New York University
Email: ako274@nyu.edu

Farhad Shirani
Department of Electrical and Computer Engineering
New York University
Email: fsc265@nyu.edu

Elza Erkip
Department of Electrical and Computer Engineering
New York University
Email: elza@nyu.edu

Yonina C. Eldar
Department of Mathematics and Computer Science
Weizmann Institute of Science
Email: yonina.eldar@weizmann.ac.il

***Abstract*—Analog-to-digital converters (ADCs) are a major contributor to the power consumption of multiple-input multiple-output (MIMO) communication systems with large number of antennas. Use of low resolution ADCs has been proposed as a means to decrease power consumption in MIMO receivers. However, reducing the ADC resolution leads to performance loss in terms of achievable transmission rates. In order to mitigate the rate-loss, the receiver can perform analog processing of the received signals before quantization. Prior works consider one-shot analog processing where at each channel-use, analog linear combinations of the received signals are fed to a set of one-bit threshold ADCs. In this paper, a receiver architecture is proposed which uses a sequence of delay elements to allow for blockwise linear combining of the received analog signals. In the high signal to noise ratio regime, it is shown that the proposed architecture achieves the maximum achievable transmission rate given a fixed number of one-bit ADCs. Furthermore, a tradeoff between transmission rate and the number of delay elements is identified which quantifies the increase in maximum achievable rate as the number of delay elements is increased.**

I. Introduction

One of the most significant challenges in the development of 5G cellular communication technologies is energy consumption. The use of large antenna arrays leads to energy demands which are inconsistent with the limited power budget available in mobile devices and small-cell access points [1]. Analog to digital converters (ADCs) are a major contributor to the power consumption in multiple-input multiple-output (MIMO) receivers. In conventional MIMO systems with digital beamforming, it is assumed that each receiver antenna is connected to a high resolution ADC [2]. In standard ADC design, the power consumption is proportional to the number of quantization bins and hence grows exponentially in the number of output bits [3], [4]. One method which has been proposed to address high power consumption in MIMO systems with large number of antennas is to use low resolution ADCs (e.g. one-bit threshold ADCs) at each receiver antenna [5]–[12]. Reducing the ADC resolution decreases power consumption, however, it also results in lower transmission rates. This suggests a tradeoff between transmission rate and power consumption which is controlled by the number and resolution of the ADCs at the receiver.

This work is supported by National Science Foundation grant SpecEES-1824434 and NYU WIRELESS Industrial Affiliates.

In classical information theory, it is well-known that in order to achieve optimal transmission rates, communication must be performed over asymptotically large blocks of data [13]. More precisely, an optimal decoder performs a possibly non-linear operation on an asymptotically large block of channel outputs. In MIMO systems using high resolution ADCs, the discretization loss is negligible due to the fine quantization grid. Simultaneous blockwise decoding is made possible by storing the digital output and performing the decoding operation over large blocklengths in the digital domain. However, when low resolution ADCs are used, discretizing the individual channel outputs prior to blockwise decoding leads to loss of information and suboptimal performance [6]. In particular, restricting to one-bit ADCs leads to large quantization noise, and a significant reduction in achievable rates [8], [10].

Rate-loss due to low resolution quantization can be attributed to two constituents which we call *intrinsic* and *extrinsic* rate-loss. To elaborate, consider the MIMO communication system shown in Fig. 1. Assume that the receiver is equipped with n_q one-bit threshold ADCs. An upper-bound on the channel capacity is given by $\min(n_q, C)$ bits per channel use, where C is the capacity of the MIMO channel when using ADCs with very high resolution. In other words, due to the restriction on the number of ADCs, the channel capacity is decreased by at least $C - \min(n_q, C)$ bits per channel-use. This *intrinsic* rate-loss cannot be reduced by improving the receiver architecture design without the use of additional one-bit ADCs. Considering the receiver architecture in Fig. 1(b), in practice only a limited set of analog operations illustrated as $f_a(\cdot)$ in the figure may be implemented. Prior works have studied the use of one-shot analog linear combiners and threshold ADCs [8]–[12]. It has been shown that the maximum rate achievable using the architecture in Fig. 1 is less than $\min(n_q, C)$ due to practical limitations in analog processing [8]. More precisely, the communication system suffers an additional *extrinsic* rate-loss of $\min(n_q, C) - R^*$ bits per channel-use, where R^* is the maximum achievable rate when these practical limitations are taken into account. In theory, the extrinsic rate-loss may be reduced by improving the receiver architecture design.

In this work, we consider communication over MIMO channels where one-bit threshold ADCs are used at the receiver. We propose a blockwise analog processing module in which

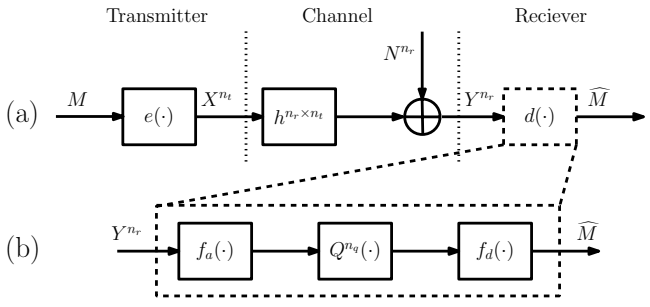

Fig. 1. The top figure shows a MIMO channel with n_t transmit antennas and n_r receive antennas. The bottom figure is the receiver architecture consisting of analog processing module $f_a(\cdot)$, n_q one-bit ADCs $Q^{n_q}(\cdot)$, and a digital processing module $f_d(\cdot)$. The function $f_a(\cdot)$ may have causal memory.

delay elements are used to reduce the extrinsic rate-loss due to one-bit ADCs. We show that for a large class of MIMO channels, in the high signal to noise ratio (SNR) regime, the proposed architecture completely eliminates the extrinsic rate-loss and achieves the maximum transmission rate among all receiver architectures with a fixed number of one-bit ADCs. We show the existence of a fundamental tradeoff between the number of delay elements and the maximum achievable rate for the proposed architecture. In addition, we show that given a fixed number of one-bit ADCs and delay elements, non-zero thresholds are necessary to achieve optimal transmission rates; whereas, using asymptotically large numbers of delay elements leads to optimal rates without requiring non-zero thresholds.

In the receiver architecture proposed in this paper, the ADC thresholds are chosen according to the channel gain matrix and are assumed to be fixed throughout the transmission block. In a companion paper [14], we propose a class of adaptive threshold receiver architectures, where the quantization thresholds at each channel-use are dependent on the channel outputs in the previous channel-uses. While the fixed threshold architecture in this paper is simpler to implement, the adaptive threshold architecture in [14] is more amiable to analysis for point-to-point (PtP) communication in the low SNR regime and multiterminal communications.

The rest of the paper is organized as follows. Section II describes the system model. Section III includes the proposed receiver architecture along with an analysis of the resulting achievable rate region. Section IV concludes the paper.

Notation: The random variable $\mathbb{1}_{\mathcal{E}}$ is the indicator function of the event $\mathcal{E}$. The set of numbers $\{1, 2, \cdots, n\}, n \in \mathbb{N}$ is represented by $[n]$. For a given $n \in \mathbb{N}$, the n-length vector $(x_1, x_2, \ldots, x_n)$ is written as x^n. The subvector $(x_k, x_{k+1}, \cdots, x_n)$ is denoted by x_k^n. We write $\|x^n\|_2$ to denote the L_2-norm of x^n. An $n \times m$ matrix is written as $h^{n\times m} = [h_{i,j}]_{i,j\in[n]\times[m]}$, and $[h^{n\times m}]^\dagger$ is the transpose of $h^{n\times m}$. Let $h_{(i)}^{n_t}, i \in [m]$ be a sequence of column vectors; the notation $[h_{(1)}^n, h_{(2)}^n, \cdots, h_{(m)}^n]^\dagger$ represents the column vector of length mn consisting of the concatenation of the original vectors. The $n \times n$ identity matrix is shown by I_n. We write $a^{n\times m} \otimes b^{n'\times m'}$ to denote the Kronecker product of matrices. The value of i modulo k is represented by $mod_k(i), i, k \in \mathbb{N}$. The binary entropy function is $h_b(x) = -x \log x - (1-x)\log(1-x)$.

II. System Model and Preliminaries

A. System Model

We consider a PtP communication system characterized by the triple $(n_t, n_r, h^{n_r\times n_t})$, where n_t is the number of transmitter antennas, n_r is the number of receiver antennas, and $h^{n_r\times n_t}$ is the channel gain matrix. The matrix $h^{n_r\times n_t}$ is assumed to be fixed over the transmission block, and known at the transmitter and receiver. The channel input and output vector pair (X^{n_t}, Y^{n_r}) is related through

$$Y^{n_r} = h^{n_r\times n_t} X^{n_t} + N^{n_r}, \tag{1}$$

where N^{n_r} is a vector of independent and identically distributed Gaussian variables with unit variance and zero mean, and the channel input has average power constraint P. It is assumed that n_q one-bit threshold ADCs are available at the receiver. The receiver uses the architecture shown in Fig. 1(b) which consists of an analog signal processing step prior to quantization and a digital signal processing step afterwards. The channel output is processed in the analog domain and the resulting vector is input to the ADCs. The output of the ADCs is processed in the digital domain to reconstruct the message. In its most general form, the analog processor may have causal memory. More precisely, the output of $f_a(\cdot)$ at time i, may depend on the matrix of received channel outputs $Y^{i\times n_r}$, where the jth row of $Y^{i\times n_r}$ is the channel output at time $j, j \le i$. Let $n \in \mathbb{N}$ be the length of the transmission block and define $\mathcal{G}_a = \{f_a : \mathbb{R}^{n\times n_r} \to \mathbb{R}^{nn_q}\}$ as the space of all functions with causal memory. Due to practical considerations, only a subset of the functions in $\mathcal{G}_a$ are implementable. We denote the space of implementable functions by $\mathcal{F}_a$. The set of implementable functions $\mathcal{F}_a$ which are considered in this paper will be discussed in Section III. The communication problem is formalized below.

Definition 1 (QMIMO). *A PtP MIMO system with one-bit ADCs (QMIMO) is characterized by the tuple $(n_t, n_r, h^{n_r\times n_t}, n_q, \mathcal{F}_a)$, where n_q is the number of one-bit ADCs, and $\mathcal{F}_a \subseteq \mathcal{G}_a$.*

Let $n, \Theta \in \mathbb{N}$ be a pair of natural numbers, $f_a \in \mathcal{F}_a$ an implementable analog function and $t^{nn_q} \in \mathbb{R}^{nn_q}$ a vector of quantization thresholds, where $t_{in_q-n_q+1}^{in_q}, i \in [n]$ is the threshold vector in the ith channel use. An $(n, \Theta, f_a, t^{nn_q})$-transmission system consists of a pair of encoding and decoding functions (e, d) where $X^{n\times n_t} = e(M)$ is the channel input over n channel uses and $\widehat{M} = d(Y^{n\times n_r}) = f_d(Q^{nn_q}(f_a(Y^{n\times n_r}) + t^{nn_q}))$ is the message reconstruction, $f_d : \{0,1\}^{nn_q} \to [\Theta]$ is a vector of Boolean functions, $Q^{nn_q}(x^{nn_q}) = (\mathbb{1}_{\{x_1\ge 0\}}, \mathbb{1}_{\{x_2\ge 0\}}, \cdots, \mathbb{1}_{\{x_{nn_q}\ge 0\}}), x^{nn_q} \in \mathbb{R}^{nn_q}$ is a sequence of one-bit ADCs. Achievability is defined in the standard Shannon sense. The capacity maximized over all implementable analog functions is denoted by $C_Q(h^{n_r\times n_t}, n_q, \mathcal{F}_a)$.

In [8], a receiver architecture is considered where the analog processing module consists of linear combiners along with non-zero threshold ADCs. The architecture is shown in Fig.

2. The linear combiner matrix $v^{n_q \times n_r}$ is applied to the received signals at each channel-use. This receiver architecture does not allow for temporal processing of the received signals in the analog domain. More precisely, the set $\mathcal{F}_a$ considered in [8] consists of all memoryless and linear analog processors:

$$\mathcal{F}_a = \{f_a |\ f_a(Y^{n\times n_r}) = (v^{n_q\times n_r} \otimes I_n)\widetilde{Y}^{nn_r}\},$$

where $\widetilde{Y}_{kn+1}^{(k+1)n}, k \in \{0, 1, \cdots, n_r - 1\}$ is equal to the kth row of $Y^{n\times n_r}$. We call this receiver architecture *one-shot*. The channel capacity using this one-shot architecture maximized over all input distributions, threshold vectors and linear combining matrices is denoted by $C_{OS}(h^{n_r\times n_t}, n_q)$.

It is known that one-shot processing of the analog signals leads to a significant extrinsic rate-loss [8], [10]. In fact, the one-shot capacity is shown to grow at most logarithmically in the number of one-bit ADCs. Consequently, in the high SNR regime, the extrinsic rate-loss due to the application of one-shot receiver architectures is at least[1] $n_q - O(\log n_q)$ and becomes arbitrarily large as the number of ADCs is increased.

In Section III, we introduce blockwise analog processing architectures which use delay elements to allow for temporal processing of the analog signals before quantization. We show in Theorem 1 that this allows us to completely eliminate the extrinsic rate-loss in a large class of MIMO systems in the high SNR regime including MIMO systems where the number of one-bit ADCs is at least twice the number of transmitter antennas and receiver antennas. To analyze the performance of the proposed architectures, we utilize a geometric interpretation introduced in [8], [10]. The geometric interpretation is especially helpful in analyzing the set of achievable rates in the high SNR regime.

The geometric interpretation and combinatorial background is briefly described in the next subsection.

B. Combinatorial Background

Loosely speaking, as the SNR is increased, the effect of noise in Equation (1) becomes negligible and the channel output is almost equal to $h^{n_r\times n_t}X^{n_t}$. In fact, in the absence of noise, the channel output space is $Im(h^{n_r\times n_t})$ the image of the channel gain matrix $h^{n_r\times n_t}$. In the following, we describe the relation between partitions of the subspace $Im(h^{n_r\times n_t})$ and the maximum transmission rate when one-shot receiver architectures are used.

Consider the one-shot architecture described in Fig. 2. For a given channel output vector y^{n_r}, let $j_i = Q(w_i), i \in [n_q]$, where $w^{n_q} = v^{n_q\times n_r}y^{n_r} + t^{n_q}$ is the input vector to the one-bit ADCs. The binary vector j^{n_q} is the vector of ADC outputs. The set of all channel output vectors y^{n_r} which result in the ADC output vector j^{n_q} is

$$\mathcal{B}_{j_1,j_2,\cdots,j_{n_q}} = \{y^{n_r} \in Im(h^{n_r\times n_t}) | Q(w_i) = j_i, i \in [n_q]\}.$$

For a given pair $(t^{n_q}, v^{n_q\times n_r})$, the collection of sets $\mathsf{B}(t^{n_q}, v^{n_q\times n_r}) = \{\mathcal{B}_{j_1,j_2,\cdots,j_{n_q}} | j_i \in \{0, 1\}, i \in [n_q]\}$, is a partition of $Im(h^{n_r\times n_t})$. The number of non-empty partition elements

[1] We write $f(x) = O(g(x))$ if $\lim_{x\to\infty} \frac{f(x)}{g(x)} < \infty$.

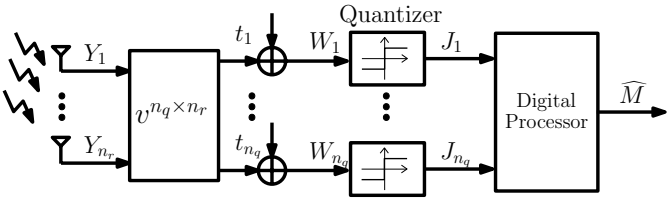

Fig. 2. A one-shot receiver architecture, where the linear combiner is characterized by the matrix $v^{n_r\times n_q}$, and the ADC thresholds are $t^{n_q} = (t_1, t_2, \cdots, t_{n_q})$.

corresponds to the number of messages which can be transmitted reliably as the SNR is taken to be asymptotically large. Note that for some binary vectors j^{n_q}, the set $\mathcal{B}_{j_1,j_2,\cdots,j_{n_q}}$ may be empty. For instance, let $n_t = n_r = 1$, $n_q = 2$, $h^{n_r\times n_t} = 1$, $v^{n_q\times n_t} = \begin{bmatrix}1\\1\end{bmatrix}$, and $t^{n_q} = \begin{bmatrix}0\\0\end{bmatrix}$. Then $\mathcal{B}_{0,1} = \{y | Q(y) = 0, Q(y) = 1\} = \varnothing$, similarly, $\mathcal{B}_{1,0} = \varnothing$. As a result, the number of partition elements may be less than 2^{n_q}. In order to increase the transmission rate, it is desirable to choose $(t^{n_q}, v^{n_q\times n_r})$ such that the number of non-empty partition elements is maximized.

We use the following proposition throughout the paper.

Proposition 1. *([15]) The maximum number of non-empty partition elements is given by*

$$\max_{t^{n_q}, v^{n_q\times n_r}} |\mathsf{B}(t^{n_q}, v^{n_q\times n_r}) - \{\varnothing\}| = \sum_{i=0}^{rank(h^{n_r\times n_t})} \binom{n_q}{i}. \tag{2}$$

Additionally , if the threshold vector is taken to be the all-zero vector, then:

$$\max_{v^{n_q\times n_r}} |\mathsf{B}(0^{n_q}, v^{n_q\times n_r}) - \{\varnothing\}| = 2 \sum_{i=0}^{rank(h^{n_r\times n_t})-1} \binom{n_q - 1}{i}. \tag{3}$$

The maximum number of non-empty partition regions grows exponentially in n_q since $\log\binom{n}{k} = nh_b(\frac{k}{n}) + O(\log n)$ [13].

III. Blockwise Receiver Architectures

We propose *blockwise* receiver architectures in which delay elements are used to perform blockwise temporal processing of the received signals before quantization. Communication is performed in $n = \ell b$ channel-uses, where n, ℓ and b are called the *blocklength*, *inner blocklength*, and *outer blocklength*, respectively. The blockwise receiver architecture uses a delay network consisting of 2ℓ delay elements as shown in Fig. 3, where each delay element $D_{n_r}(\cdot)$ takes the vector of received signals at the ith channel-use $Y^{n_r}(i)$ and outputs $Y^{n_r}(i-1)$. In other words, $D_{n_r}(\cdot)$ delays the received analog vector by one channel-use. The stored analog signals are combined using the linear combining matrix $v^{\ell n_q\times \ell n_r}$ over ℓ channel-uses.

To clarify the linear combination process, let us describe the first 3ℓ channel-uses. In the first ℓ channel-uses the received signals $Y^{n_r}(i), i \in [\ell]$ are stored in the delay network. In the next ℓ channel-uses, the second batch of received signals $Y^{n_r}(i), \ell + 1 \le i \le 2\ell$ are stored in the delay network while the linear combiner operates on the previously stored signals $Y^{n_r}(i), i \in [\ell]$. More precisely, for the ith channel-use where

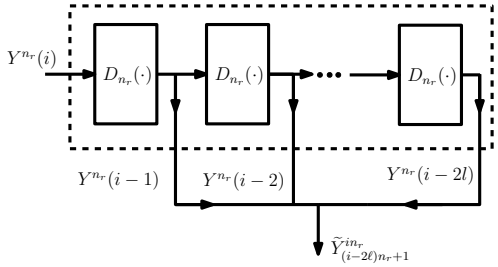

Fig. 3. Delay network in the ith channel-use where $\widetilde{Y}^{in_r}_{(i-2\ell)n_r+1} = (Y^{n_r}(i-2\ell+1), Y^{n_r}(i-2\ell+2), \cdots, Y^{n_r}(i-1), Y^{n_r}(i))$ and $i \geq 2\ell$.

$\ell+1 \leq i \leq 2\ell$, the linear combiner outputs $\overline{Y}^{(i-\ell)n_q}_{(i-\ell-1)n_q+1}$, where $\overline{Y}^{\ell n_q} = v^{\ell n_q \times \ell n_r} \widetilde{Y}^{\ell n_r}$, and

$$\widetilde{Y}^{\ell n_r} = (Y^{n_r}(1), Y^{n_r}(2), \cdots, Y^{n_r}(\ell-1), Y^{n_r}(\ell)).$$

In the third ℓ channel-uses, the third batch of received signals $Y^{n_r}(i), 2\ell+1 \leq i \leq 3\ell$ are stored in the delay network while the linear combiner operates on the previously stored signals $Y^{n_r}(i), i \in [\ell]$. This process continues for b blocks of length ℓ until the nth channel-use, where n is the blocklength. The output of the linear combiner is given to the n_q one-bit threshold ADCs. The threshold vector used in the one-bit ADCs changes periodically with a period of ℓ channel-uses. More precisely, let $t^{\ell n_q} \in \mathbb{R}^{\ell n_q}$ and define $\widetilde{t}^{n_q}(k) = t^{kn_q}_{kn_q-n_q+1}, k \in [\ell]$. For the first ℓ channel-uses, the threshold vector $\widetilde{t}^{n_q}(i)$ is used in the ith channel-use. For the second ℓ channel-uses, the threshold vector $\widetilde{t}^{n_q}(i-\ell)$ is used in the ith channel-use. Generally, for $i \in [n]$, let $k = mod_\ell(i)$, the vector $t^{n_q}(k)$ is used as the threshold vector for the one-bit ADCs at the ith channel-use.

We call the resulting communication system a D-QMIMO system, where D refers to delay. The set of implementable analog functions for this architecture is:

$$\mathcal{F}_a^{\ell} = \{f_a |\ f_a(Y^{n\times n_r}) = (v^{\ell n_q \times \ell n_r} \otimes I_b)\widetilde{Y}^{nn_r}, b \in \mathbb{N}\},$$

where

$$\widetilde{Y}^{jn_r}_{(j-2\ell)n_r+1} = (Y^{n_r}(j-2\ell+1), Y^{n_r}(j-2\ell+2), \cdots, Y^{n_r}(j-1), Y^{n_r}(j)),$$

where $2\ell \leq j \leq n$. The channel capacity optimized over all analog combining matrices, and threshold vectors is denoted by $C_\ell(h^{n_r\times n_t}, n_q)$ for a given delay ℓ.

Note that D-QMIMO systems are a special class of QMIMO systems where the analog processing is restricted to linear operations. Furthermore, the one-shot setup described in Fig. 2 is a special case of D-QMIMO where the length of each inner-block is equal to one (i.e. $\ell = 1$). As a result,

$$C_{OS}(h^{n_r\times n_t}, n_q) = C_1(h^{n_r\times n_t}, n_q) \leq C_\ell(h^{n_r\times n_t}, n_q),$$

where $\ell \geq 2$.

We derive the bounds provided in Theorem 1 below on the performance of the following proposed coding strategy. Consider a D-QMIMO communication system where $(t^{\ell n_q}, v^{\ell n_q \times \ell n_r})$ are taken so that the partition $\mathsf{B}(t^{\ell n_q}, v^{\ell n_q \times \ell n_r})$ has the maximum number of non-empty elements as described in Proposition 1. In other words, $(t^{\ell n_q}, v^{\ell n_q \times \ell n_r})$ are chosen such that the number of non-empty partitions is equal to $\sum_{i=0}^{\ell rank(h^{n_r\times n_t})} \binom{\ell n_q}{i}$. We use the fact that $\log\binom{n}{k} = nh_b(\frac{k}{n}) + O(\log n)$ [13] and perform a second order analysis of the number of non-empty partition elements as the number of delay elements ℓ is increased asymptotically to characterize the set of achievable rates for high SNRs.

Theorem 1. *For the D-QMIMO communication system with n_q one-bit ADCs, the capacity C_ℓ satisfies the following*

$$n_q h_b(\alpha) - \frac{\log(\ell)}{2\ell} + O\left(\frac{1}{\ell}\right) \leq C_\ell \leq n_q h_b(\beta) - \frac{\log(\ell)}{2\ell} + O\left(\frac{1}{\ell}\right),$$

as SNR$\to\infty$, *where* $\alpha = \min\{\frac{rank(h^{n_r\times n_t})}{n_q}, \frac{1}{2}\}$, $\beta = \min\{\frac{n_r}{n_q}, \frac{1}{2}\}$. *Particularly, if* $rank(h^{n_r\times n_t}) = n_r$, *then*

$$C_\ell \to n_q h_b(\beta) \text{ as } \ell \to \infty.$$

An outline of the proof is provided in the Appendix, where a general coding strategy for arbitrary SNRs is presented. The resulting rate is analyzed in the high SNR regime. The complete proof is given in [16].

The following observations follow from Theorem 1:

I) The capacity approaches n_q as $\ell \to \infty$ if $n_q \leq 2rank(h^{n_r\times n_t})$. Consequently, the extrinsic rate-loss is completely eliminated. This is in contrast with prior works (e.g. [8], [10]), where the high SNR capacity grows logarithmically in n_q.

II) The maximum achievable rate due to using non-zero threshold ADCs is $\frac{1}{\ell}\log\sum_{i=0}^{\ell rank(h^{n_r\times n_t})}\binom{\ell n_q}{i}$, whereas when zero threshold ADCs are used, the maximum rate is $\frac{1}{\ell}\log\sum_{i=0}^{\ell rank(h^{n_r\times n_t})} 2\binom{\ell n_q-1}{i}$. The two values converge to each other as $\ell \to \infty$. This shows that when long delays ℓ can be tolerated, zero threshold ADCs can be used in scenarios where non-zero thresholds are costly to implement without any loss in transmission rate.

III) For a fixed number of transmitters n_t and receivers n_r, as the number of one-bit ADCs n_q is increased, the maximum achievable rate increases linearly when $n_q \leq 2rank(h^{n_r\times n_t})$ since $h_b(\alpha) = h_b(\frac{1}{2}) = 1$. The maximum achievable rate increases logarithmically when $n_q \gg 2rank(h^{n_r\times n_t})$ since

$$n_q h_b\left(\frac{rank(h^{n_r\times n_t})}{n_q}\right) = rank(h^{n_r\times n_t})(\log n_q - O(\log n_q)).$$

This is shown in Fig. 4, where for a MIMO system with $n_r = 10$, the achievable rate in Theorem 1 is plotted as a function of n_q for $n_t \in \{2,4,6,8\}$ as the number of delay elements is taken to be asymptotically large.

IV. Conclusion

We have considered point-to-point communication over MIMO systems when a limited number of one-bit ADCs are available at the receiver. We have proposed a receiver architecture which uses a sequence of delay elements to allow for blockwise linear combining of the received analog signals. In the high SNR regime, given a fixed number of one-bit ADCs, we have shown that the proposed architecture achieves the maximum transmission rate among all receiver architectures. Furthermore, we have characterized a tradeoff between

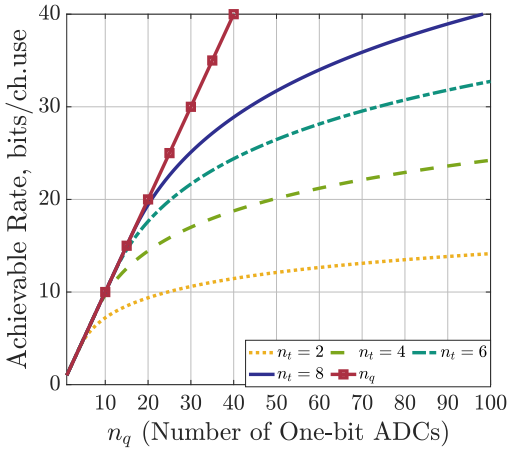

Fig. 4. The figure shows the maximum achievable high SNR rate when the number of delay elements ℓ is taken to be asymptotically large for the MIMO system with $n_r = 10$ and $n_t \in \{2, 4, 6, 8\}$. The red full line is the $R = n_q$ line which is achievable if n_t, n_r are asymptotically large. The channel is assumed to be full-rank.

transmission rate and the number of delay elements which quantifies the increase in maximum achievable rate as the number of delay elements is increased. In a companion paper [14] we propose a class of adaptive threshold architectures analyze their performance in PtP communications in the low SNR regime and broadcast channel communications.

References

[1] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2014.

[2] S. Han, I. Chih-Lin, Z. Xu, and C. Rowell, "Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5g," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, 2015.

[3] R. H. Walden, "Analog-to-digital converter survey and analysis," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 4, pp. 539–550, 1999.

[4] Y. C. Eldar, *Sampling Theory: Beyond Bandlimited Systems*. Cambridge University Press, 2015.

[5] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath, "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in *IEEE 48th Asilomar Conference on Signals, Systems and Computers*, 2014, pp. 957–961.

[6] A. Mezghani and J. A. Nossek, "Capacity lower bound of MIMO channels with output quantization and correlated noise," in *IEEE International Symposium on Information Theory (ISIT)*, 2012.

[7] O. Orhan, E. Erkip, and S. Rangan, "Low power analog-to-digital conversion in millimeter wave systems: Impact of resolution and bandwidth on performance," in *Information Theory and Applications Workshop (ITA)*, 2015, pp. 191–198.

[8] A. Khalili, S. Rini, L. Barletta, E. Erkip, and Y. C. Eldar, "On MIMO channel capacity with output quantization constraints," in *IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 1355–1359.

[9] S. Rini, L. Barletta, Y. C. Eldar, and E. Erkip, "A general framework for MIMO receivers with low-resolution quantization," *IEEE Information Theory Workshop*, Nov. 2017.

[10] J. Mo and R. W. Heath, "Capacity analysis of one-bit quantized MIMO systems with transmitter channel state information," *IEEE Transactions on Signal Processing*, vol. 63, no. 20, pp. 5498–5512, 2015.

[11] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 122–131, 2014.

[12] T. Koch and A. Lapidoth, "At low SNR, asymmetric quantizers are better," *IEEE Transactions on Information Theory*, vol. 59, no. 9, pp. 5421–5445, 2013.

[13] I. Csiszar and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[14] A. Khalili, F. Shirani, E. Erkip, and Y. C. Eldar, "On multiterminal communication over MIMO channels with one-bit ADCs at the receivers," *to appear in IEEE International Symposium on Information Theory (ISIT)*, 2019.

[15] R. Winder, "Partitions of n-space by hyperplanes," *SIAM Journal on Applied Mathematics*, vol. 14, no. 4, pp. 811–818, 1966.

[16] A. Khalili, F. Shirani, E. Erkip, and Y. C. Eldar, "Tradeoff between delay and high SNR capacity in quantized MIMO systems," *arXiv preprint arXiv.org*, 2019.

Proof of Theorem 1

To prove the achievability (lower bound on C_ℓ), we describe an outline of the coding strategy for arbitrary SNRs, where the average transmission power constraint is $\mathbb{E}(\|X^{n_t}\|_2^2) \leq P$. The resulting communication rate is then analyzed as SNR$\to \infty$. Fix ℓ and b, where b is the outer code blocklength.

Consider the pair $(t^{\ell n_q}, v^{\ell n_q \times \ell n_r})$ which achieve the maximum number of non-empty sets in Proposition 1. For a given rate R, define $\Theta_n = 2^{nR}$, where $n = b\ell$. The message $M \in [\Theta_n]$ is transmitted over $b+1$ transmission blocks each of ℓ symbols, where each symbol is in $\mathbb{R}^{n_t}$. Let $\mathsf{B}(t^{\ell n_q}, v^{\ell n_q \times \ell n_r}) = \{\mathcal{B}_{j^{\ell n_q}} | j_i \in \{0,1\}, i \in [\ell n_q]\}$ be the partition corresponding to the pair $(v^{\ell n_q \times \ell n_r}, t^{\ell n_q})$ as defined in Section II. Define $\mathcal{J} = \{j^{\ell n_q} \in \{0,1\}^{\ell n_q} | \mathcal{B}_{j^{n_q}} \neq \emptyset\}$, and let $\hat{y}^{\ell n_r}_{j^{\ell n_q}} \in \mathcal{B}_{j^{\ell n_q}}$, $j^{\ell n_q} \in \mathcal{J}$ be a set of representatives for the partition elements. We define the input vector corresponding to $\hat{y}^{\ell n_r}_{j^{\ell n_q}}$ as

$$\hat{x}^{\ell n_t}_{j^{\ell n_q}} = argmin_{\hat{y}^{\ell n_r}_{j^{\ell n_q}} = (h^{n_r \times n_t} \otimes I_\ell) x^{\ell n_t}} \|x^{\ell n_t}\|_2, \tag{4}$$

and the cost associated with $\hat{y}^{\ell n_r}_{j^{\ell n_q}}$ as:

$$c(\hat{y}^{\ell n_r}_{j^{\ell n_q}}) = min_{\hat{y}^{\ell n_r}_{j^{\ell n_q}} = (h^{n_r \times n_t} \otimes I_\ell) x^{\ell n_t}} \|x^{\ell n_t}\|_2.$$

We define the random vector $Z^{\ell n_r}$ through the transition probability $P_{Z^{\ell n_r}|\widehat{Y}^{\ell n_t}}$, where:

$$P_{Z^{\ell n_r}|\widehat{Y}^{\ell n_t}}(\hat{y}^{\ell n_r}_{k^{\ell n_q}} | \hat{y}^{\ell n_r}_{j^{\ell n_q}}) = P\left(v^{\ell n_q \times \ell n_r}(\hat{y}^{\ell n_r}_{j^{\ell n_q}} + N^{\ell n_r}) + t^{\ell n_q} \in \mathcal{B}_{k^{\ell n_q}}\right),$$

where $j^{\ell n_q}$, $k^{\ell n_q} \in \mathcal{J}$. Let C_{outer} be the capacity of the PtP channel with the transition probability $P_{Z^{\ell n_r}|\widehat{Y}^{\ell n_r}}$ subject to the average power constraint $\mathbb{E}(c(\widehat{Y}^{\ell n_r})) \leq \ell P$. We first construct a family of capacity achieving codes for this channel using standard random coding methods. Each symbol $\widehat{Y}^{\ell n_r}_{j^{\ell n_q}}$ in the randomly generated codewords has alphabet $\mathbb{R}^{\ell n_r}$. In order to transmit the symbol $\widehat{Y}^{\ell n_r}_{j^{\ell n_q}}$, the transmitter finds the corresponding input $X^{\ell n_t}$ (Equation (4)) and transmits the vector over ℓ channel uses. As the SNR goes to infinity, the channel $P_{Z^{\ell n_r}|\widehat{Y}^{\ell n_t}}$ becomes noiseless. It can be shown that the capacity is equal to $\frac{1}{\ell} H(\widehat{Y}^{\ell n_r}) = \frac{1}{\ell} \log \sum_{i=0}^{\ell rank(h^{n_r \times n_t})} \binom{\ell n_q}{i}$. We use the fact that $\log \binom{n}{k} = nh_b(\frac{k}{n}) + O(\log n)$ to show that the expression converges to the achievable rate in Theorem 1. The converse follows by using the Fano's inequality along with Proposition 1 and is omitted due to space limitations. The complete proof is given in [16].