New Equivalent Model of a Quantizer With Noisy Input and Its Applications for MIMO System Analysis and Design

ARKADY MOLEV-SHTEIMAN\textsuperscript{1}, (Member, IEEE), XIAO-FENG QI\textsuperscript{1}, (Senior Member, IEEE), LAURENCE MAILANDER\textsuperscript{1}, (Member, IEEE), NARAYAN PRASAD\textsuperscript{1}, AND BERTRAND M. HOCHWALD\textsuperscript{2}, (Fellow, IEEE)

\textsuperscript{1}Radio Algorithms Research, Futurewei Technologies, Inc., Bridgewater, NJ 08807, USA
\textsuperscript{2}Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

Corresponding author: Arkady Molev-Shteiman (arkady.molev.shteiman@futurewei.com)

\section*{ABSTRACT} We propose a novel equivalent model for a quantizer with noisy input (the desired signal corrupted by measurement noise). It presents the quantizer output as a sum of the desired signal after it passes through a nonlinear element with a known equivalent transfer function and an equivalent additive white noise. The equivalent transfer function takes the form of a conditional expectation of the quantizer output given the desired signal portion of its input. The proposed model proves to be effective for the analysis and design of MIMO systems employing low-resolution quantizers (analog to digital and digital to analog converters, ADCs and DACs, respectively). We also demonstrate the efficacy of the model through several example applications for 1) the design of digital dither that mitigates the effect of DAC quantization error in a MIMO transmitter and significantly reduces the DAC resolution requirement; 2) the determination of the minimal ADC resolution required for operation of conventional MIMO receivers designed for infinite-resolution ADC arrays, without incurring significant performance degradation; and 3) the design of simple MIMO receivers (ML and MMSE) that mitigate the effect of insufficient ADC resolution, thereby extending the receiver SNR operating range without an undue complexity increase.

\section*{INDEX TERMS} Massive MIMO, low-resolution ADC, low-resolution DAC, Bussgang-Rowe decomposition.

\section*{I. INTRODUCTION} Massive MIMO is an emerging technology capable of improving the spectral efficiency of wireless communication by orders of magnitude \cite{1}. However, a significant increase in base station antennas implies a proportional increase in cost and power consumption. On the other hand, it was shown that massive MIMO may significantly mitigate the impact of imperfections in the hardware implementation \cite{2}, implying that we may use less expensive and lower energy components to implement massive MIMO. In the overall cost and energy budget of massive MIMO, base station quantizers (analog to digital converters - ADCs and digital to analog converters DACs) are important elements. It is known that the power consumption of the quantizer grows exponentially with the number of quantization bits \cite{3}. Therefore, algorithms that reduce quantizer resolution have significant practical importance.

Intuitively, the quantizer resolution should be sufficiently high to ensure that quantization noise power is sufficiently below that of thermal noise. As the number of antennas in a MIMO system scales, we can operate at lower per antenna SNR. In the latter regime, the thermal noise power at ADC input grows. It follows that the tolerable quantization error increases and quantizer resolution decreases accordingly, to a single bit in the extreme.

The conventional approach assumes that ADC resolution is sufficiently high, so we can approximate the quantization error as an independent white process with variance $\Delta^2/12$, where $\Delta$ is the quantization step \cite{4}. In such cases, we can set the ADC resolution to ensure that quantization noise variance is much lower than that of thermal noise variance,
so we can neglect it. However, for low-resolution ADCs, this approximation does not work, and accurate analysis of ADC must be done.

The problem of low-resolution AD/DA in the context of multi-antenna massive MIMO has attracted heavy research interests lately. Many contributions that consider uplink massive MIMO receivers with arrays of low-resolution ADCs have been published. A general overview of low-resolution ADC is given in [9]. An information-theoretic analysis of massive MIMO uplink receivers with low-resolution ADC performance is given in [10]–[30]. Other works propose different methods of data reception [31]–[43] and channel estimation [40], [44]–[51], taking into account limited ADC resolution. Lately, MIMO radar with low-resolution ADC has attracted research attention [54], [55]. Separately, massive MIMO transmitters employing low-resolution DACs have emerged as an active area of investigation. Different methods of low-resolution DAC precoding were presented in [53]–[62].

We identify two major accepted approaches of dealing with finite-resolution ADCs.

- The linear approach that presents the ADC output as the sum of a scaled version of the total input signal and a nonlinear distortion (NLD) that is uncorrelated with the input signal. There are two methods to obtain this decomposition:
  - The additive quantization noise model (AQNM) [4] assumes that the ADC input given ADC output is equal to ADC output. It implies a special ADC designed to match the input probability density function (PDF), such as the Lloyd-Max quantizer [4]. For the commonly used uniform ADC, this is an approximation that is only valid at high resolution (see [10]–[17] and [31]–[34]).
  - The Bussgang-Rowe decomposition method [5], [6] is more general but also more complex (see [28]–[30] and [51]).

- Probabilistic methods that search for the desired input signal that maximizes the likelihood of the observed ADC output vector (see [18]–[27], [35]–[43] and [40], [44]–[49]).

Our model explicitly accounts for communication and sensing inputs that consist of a desired part and noise (i.e., noisy inputs). It represents ADC output as the sum of the desired signal alone passing through an equivalent transfer function (ETF), and a white noise component, where the ETF is given as the expectation of the quantizer output conditioned on the desired part of the input signal, taken over the noise ensemble. Here, we refer to our model as the conditional expectation model (CEM).

The ETF captures the well-known linearizing effects of additive noise (dither) shown in [64]–[66]. However, we believe that we are the first to apply it to the quantized MIMO system problem.

As needed, Bussgang-Rowe decomposition can be applied to the ETF to extract its equivalent linear gain and NLD.

Our application of the Bussgang-Rowe decomposition extracts the linear gain for only the desired signal, thus maximizing the power of the white noise portion of the equivalent output noise. Indeed, in the sequel we highlight and quantify the advantages of applying Bussgang-Rowe decomposition to the ETF as opposed to the conventional approach of applying it to the original quantizer transfer function.

We summarize our contributions as follows.

- We propose a revised quantizer equivalent model that facilitates analysis and design of systems employing finite-resolution quantizers for a practically relevant class of inputs, namely, those consisting of a desired signal corrupted by noise. Such inputs are ubiquitous in communication and sensing systems yet have not been explicitly and systematically addressed by the existing analytic models to the best of our knowledge.

- We apply the revised model to a canonical multiuser MIMO link over a line-of-sight (LoS) channel and provide an analysis of the separate and distinctive effects of the white and nonwhite components of the quantizer output error on system performance.

- The ETF dependence on noise statistics allows it to be linearized by a properly designed artificial dither. We propose a simple DAC dithering scheme based on the CEM applied to a MIMO transmitter that significantly reduces DAC resolution and/or transmit EVM seen by a receiver.

- We propose an ADC resolution determination methodology based on the CEM applied to MIMO systems, which provides a worst-case performance guarantee for a MIMO system employing naïve receivers that do not consider the effect of quantization. The incentive to reuse close-to-legacy naïve receivers designed for high-resolution ADCs has been lower hardware complexity and power as well as a shorter design cycle. Previous ADC resolution determination works [10]–[13] make the approximating assumption that quantization error is white. We provide a resolution determination methodology that analyzes worst-case realization of the nonwhite component of quantization error. As our work demonstrates, this nonwhite component strongly affects the performance of the MIMO receiver with low-resolution ADC. The SNR regime in which the performance of a naïve receiver is acceptable is predicted and in excellent agreement with that from numerical simulation. The design rules are articulated in the lexicon familiar to practical system designers, e.g., the noise figure, making them accessible to practicing engineers and complementary to information-theoretic guidelines.

- We demonstrate the efficacy and accuracy of the model by devising simple MIMO receivers that compensate for low-resolution-induced distortion, resulting in near-optimal performance. This simplification is a direct consequence of the CEM presenting instantaneous NLD as a function of signal/noise statistics, but only the desired part of the instantaneous input signal without its...
noisy part, precisely the part that needs compensation for nonlinearity.

The rest of the paper is organized as follows:
- Section II presents a new equivalent model of a quantizer with a noisy input.
- Section III presents the application of our model for the design of digital dither that mitigates the effect of quantization error in the downlink MIMO transmitter and significantly reduces requirements to its DAC resolution.
- Section IV presents the uplink MIMO receiver system model.
- Section V presents an application of our model for determination of the minimal ADC resolution required for operation of a naïve MIMO receiver that does not consider the effect of low-resolution ADC without significant performance degradation. For this purpose we tailor Bussgang-Rowe decomposition
- Section VI provides examples of an application of our model to the design of low-complexity MIMO receivers (ADCs and DACs).

II. THE CONDITIONAL EXPECTATION MODEL OF A QUANTIZER WITH NOISY INPUT

A real input quantizer (ADC or DAC) performs the quantization operation:

\[
Q_R(s) = \arg \min_{q_r} \{ |s - q_r| \} \tag{1}
\]

where \( s \in \mathbb{R} \) is the quantizer input, and \( q_r \) is the element \( r \) of the set of \( R \) possible quantizer output levels.

\[ R = 2^{N_{\text{BIT}}} \tag{2} \]

where \( N_{\text{BIT}} \) denotes the number of ADCs bits (resolution).

For a uniform quantizer output level, \( q_r \) is equal to:

\[ q_r = (2 \cdot r - R - 1) \cdot \left( \frac{\Delta}{2} \right) \quad \text{for} \quad (r = 1, 2, \ldots, R) \tag{3} \]

where \( \Delta \) is the quantization step size.

As an example, the 2-bit uniform quantizer transfer function is shown in figure 1.

Let us define the complex quantizer operation as:

\[
Q(s) = Q_R(\text{Re}(s)) + j \cdot Q_R(\text{Im}(s)) \quad \text{where:} \quad j = \sqrt{-1} \tag{4}
\]

Let us consider a quantizer that receives as an input the sum of a desired signal \( s_I(m) \) that should be passed with minimum distortion and an additive noise \( n_I(m) \). The quantizer output \( s_Q(m) \) is defined as:

\[
s_Q(m) = Q(s_I(m) + n_I(m)) \tag{5}
\]

where \( m \) is any signal index (for example, a time sample index or an antenna index).

Let us assume that the real and imaginary parts of the additive input noise \( n_I(m) \) on each quantizer are independent from each other and have an identical probability density function (PDF) \( p_N(x) \).

We do not make any assumption about the input desired signal except that its real and imaginary parts are uncorrelated and have an identical normalized PDF \( p_S(x) \) with zero mean.

Let us define the conditional expectation (over the noise ensemble) of the noisy quantizer output \( s_Q \) given quantizer input \( s_I \) as the quantizer “equivalent transfer function” \( F(s_I) \). Because real and imaginary parts of the quantizer input noise are independent:

\[
F(s_I(m)) \triangleq E[s_Q(m) | s_I(m)]
\]

\[ = F_R(\text{Re}(s_I(m))) + j \cdot F_R(\text{Im}(s_I(m))) \tag{6} \]

where \( F_R(s) \) is the convolution of the true quantizer transfer function and the noise PDF.

\[
F_R(x) \triangleq E[\text{Re}(s_Q(m)) | \text{Re}(s_I(m)) = x]
\]

\[ = E[\text{Im}(s_Q(m)) | \text{Im}(s_I(m)) = x]
\]

\[ = \int_{n=-\infty}^{\infty} Q_R(x + n) \cdot p_N(n) \cdot dn = \sum_{r=1}^{R} q_r \cdot \Pr(q_r | x) \tag{7} \]

where:

\[
\Pr(q_r | x) \triangleq \Pr(\text{Re}(s_Q(m)) = q_r | \text{Re}(s_I(m)) = x)
\]

\[ = \Pr(\text{Im}(s_Q(m)) = q_r | \text{Im}(s_I(m)) = x)
\]

\[
= \begin{cases} 
0.5 \cdot (q_r + q_{r+1})^x, & \text{if } (r = 1) \\
0.5 \cdot (q_{r-1} + q_r)^{-x}, & \text{if } (r = R) \\
\int_{n=-\infty}^{+\infty} p_N(n) \cdot dn, & \text{else} \\
\int_{n=0.5 \cdot (q_r - q_{r+1})^{-x}}^{0.5 \cdot (q_r + q_{r+1})^{-x}} p_N(n) \cdot dn & \text{if } (r, R - 1) \\
\int_{n=0.5 \cdot (q_{r-1} + q_r)^{-x}}^{0.5 \cdot (q_r - q_{r+1})^{-x}} p_N(n) \cdot dn & \text{else}
\end{cases} \tag{8} \]
Let us define the quantizer equivalent output noise as,

\[
    n_Q \triangleq (m) s_Q (m) - F (s_I (m))
\]

In Appendix A, we prove that the quantizer equivalent noise is a white process uncorrelated with the quantizer input signal. It has the following properties:

\[
    E [n_Q(m)] = 0 \quad \text{for any } m
\]

\[
    E [\text{Re} (n_Q(m)) \cdot \text{Im} (n_Q(m))] = 0 \quad \text{for any } m
\]

\[
    E [n_Q (m_1)^* \cdot s_I (m_2)] = 0 \quad \text{for any pair } m_1 \text{ and } m_2
\]

\[
    E [n_Q (m_1)^* \cdot F (s_I (m_2))] = 0 \quad \text{for any pair } m_1 \text{ and } m_2
\]

\[
    E [n_Q (m_1)^* \cdot n_Q (m_2)] = 0 \quad \text{for any pair } (m_1 \neq m_2)
\]

In Appendix A, we also express the equivalent output noise variance for the scenario when the real and imaginary parts of the desired signal \(s_I (m)\) are independent from each other and have an identical PDF \(p_S \) (x),

\[
    \sigma_Q^2 = E \left[|n_Q (m)|^2 \right] = 2 \cdot \int_{x=-\infty}^{\infty} \left(V_R (x) - F_R (x)^2\right) \cdot p_S (x) \cdot dx
\]

where the energy function \(V_R (s_I)\) is defined as,

\[
    V_R (x) \triangleq E \left[\text{Re} \left(s_Q (m)\right)^2 \mid \text{Re} \left(s_I (m)\right) = x\right] = E \left[\text{Im} \left(s_Q (m)\right)^2 \mid \text{Im} \left(s_I (m)\right) = x\right]
\]

\[
    = \int_{n=-\infty}^{\infty} Q_R (x + n)^2 \cdot p_N (n) \cdot dn = \sum_{r=0}^{K-1} q_r^2 \cdot Pr (q_r | x)
\]

The new equivalent model of the quantizer with a noisy input represents the quantizer output as the sum of the desired signal passing through the equivalent transfer function and the equivalent additive white noise at the output.

\[
    Q (s_I (m) + n_I (m)) = F (s_I (m)) + n_Q (m)
\]

The block diagram of the equivalent model for a quantizer with a noisy input is shown in figure 2.

The CEM model is applicable for any ADC type. For brevity, in the sequel, we consider only a uniform quantizer, which is of the greatest practical interest.

III. DOWNLINK MIMO TRANSMITTER DAC RESOLUTION REDUCTION

A. SYSTEM MODEL

Let us consider an ‘all digital’ downlink MIMO transmitter equipped with \(M\) antennas that serves \(K\) users with the goal to deliver to each user a communication signal \(x_k\) with variance \(\sigma^2\). Let us assume for simplicity a flat fading channel. We can always extend our results to a nonflat (frequency-selective) fading channel by adding the dimension of time to all our equations. For a flat fading channel, without considering the DAC resolution, a transmitter sends through each antenna \(m\) a signal \(s_I (m)\) such that the summed signal that arrives at each user \(k\) is equal to:

\[
    x_k = \sum_{m=1}^{M} h_k (m) \cdot s_I (m) \quad k = 1, 2, \ldots, K
\]

where \(h_k (m)\) is the downlink channel between antenna \(m\) and user \(k\).

However, due to DAC quantization, the actual signal that arrives at user \(k\) is:

\[
    \tilde{x}_k = \sum_{m=1}^{M} h_k (m) \cdot Q (s_I (m)) \quad k = 1, 2, \ldots, K
\]

Then, the DAC output can be represented as the sum of the desired signal and quantization error:

\[
    Q (s_I (m)) = s_I (m) + w_Q (m)
\]

where:

\[
    w_Q (m) \triangleq Q (s_I (m)) - s_I (m)
\]

For simplicity, this article does not consider the clipping effect of the DAC:

\[
    (q_1 \leq \text{Re} (s_I (m)) < q_2, q_1 \leq \text{Im} (s_I (m)) < q_2)
\]

We also assume that DAC resolution is sufficiently high to approximate the PDF of both the real and imaginary parts of the input signal within each quantization step \(\Delta\) as uniform.

\[
    p_S (x) \approx (1/\Delta) \cdot \Pi (|x - q_r| \leq (\Delta/2))
\]

The PDF is also uniform within an interval \(\pm 0.5 \cdot \Delta\).

\[
    p_w (w) = (1/\Delta) \cdot \Pi (w/\Delta)
\]

where:

\[
    \Pi (x) = \begin{cases} 
        1 & \text{if } (-0.5 \leq x < +0.5) \\
        0 & \text{else}
    \end{cases}
\]
Therefore, the quantization error has zero mean and variance equal to:

\[
\sigma_{WO}^2 = E \left[ w_O (m)^2 \right] = 2 \cdot \int_{-\infty}^{+\infty} w^2 \cdot p_w (w) \cdot dw
\]

\[= 2 \cdot \int_{w=-0.5\Delta}^{0.5\Delta} w^2 \cdot dw = 2 \cdot \frac{\Delta^2}{12} = \frac{\Delta^2}{6} \quad (26)\]

From (18), (19) and (20), it follows that the actual signal that arrives at user \( k \) is equal to:

\[\hat{x}_k = x_k + \sum_{m=0}^{M-1} h_k (m) \cdot w_O (m) \quad (27)\]

The quantization distortion of the signal arriving at user \( k \) is measured by error vector magnitude (EVM), here defined as the ratio between the variance of distortion (quantization error) and variance of the desired signal,

\[EVM_{O,k} \triangleq \frac{E \left[ x_k - \hat{x}_k \right]^2}{\sigma_X^2} = \frac{E \left[ \sum_{m=0}^{M-1} h_k (m) \cdot w_O (m) \right]^2}{\sigma_X^2} \quad (28)\]

Because \( w_O (m) \) is not necessarily a white process, according to the statement proven in Appendix B and (28), the worst-case EVM is equal to:

\[\text{max} \left( EVM_{O,k} \right) = \frac{\left( \sum_{m=0}^{M-1} |h_k (m)| \right)^2}{\sigma_X^2} \cdot \frac{\sigma_{WO}^2}{\Delta^2} \quad (29)\]

This worst case occurs when quantization errors of all antennas sum coherently and resonate.

In Appendix C, we proved that one example of such a worst case is the maximum ratio transmitter (MRT) for a single-user scenario with an LoS channel:

\[h_0 (m) = g_0 \cdot \exp (j \cdot \pi \cdot \sin (\alpha_0) \cdot m) \quad (30)\]

The angle of arrival is equal to:

\[\alpha_0 = 0 \quad \text{or} \quad \pm \pi/2 \text{ or } \pm \pi/3 \quad (31)\]

Each communication standard specifies the worst-case distortion level (or, maximal EVM) of the signal that arrives at a user. Based on this, we can determine the minimal DAC resolution that generates distortions below this permissible limit.

\[\text{B. DAC WITH OPTIMAL DIGITAL DITHER}\]

To prevent the worst case, we want to make the quantization error a white process across transmit antennas. To accomplish this, we have to linearize the DAC equivalent transfer function by adding to each DAC input signal \( s_t (m) \) the dithering signal \( n_I (m) \) from an independent digital random number generator.

\[s_O (m) = Q (s_t (m) + n_I (m)) \quad m = 0, 1, \ldots, M - 1 \quad (32)\]

The real and imaginary parts of the dithering signal \( n_I (m) \) are also independent and have an identical PDF \( p_N (n_I) \). The optimal dither PDF that linearizes the DAC equivalent transfer function within the dynamic range (22), according to (7) must satisfy:

\[F_R (x) = \int_{y=-\infty}^{\infty} Q_R (x + n_I) \cdot p_N (n_I) \cdot dn_I\]

\[= \sum_{r=1}^{R} q_r \cdot \Pr (q_r | x) = \begin{cases} q_1 & \text{if } (x \leq q_1) \\ q_R & \text{if } (x > q_R) \\ x & \text{else} \end{cases} \quad (33)\]

The dither PDF that satisfies condition (33) is the uniform PDF over the interval \( \pm \Delta/2 \):

\[p_N (n_I) = \frac{1}{\Delta} \cdot \Pi (n_I / \Delta) \quad (34)\]

According to (8) and (3), for such a dither PDF we have:

\[\Pr (q_r | x) = \begin{cases} (x-q_r)/\Delta & \text{if } (0 \leq (x-q_r) < (\Delta/2)) \\ (q_r-x)/\Delta & \text{if } (0 < (q_r-x) \leq (\Delta/2)) \\ 0 & \text{else} \end{cases} \quad (35)\]

Figure 3 illustrates the linearization effect as the result of convolution (7):

From (33) and (17), it follows that within the DAC dynamic range (22), we can represent its output as the sum of the desired signal and the equivalent additive noise,

\[Q (s_t (m) + n_I (m)) = F (s_t (m)) + n_O (m) = s_t (m) + n_O (m) \quad (36)\]

The equivalent additive noise \( n_O (m) \) is a white process.

According to (15), (16) and (35) for the input signal PDF given by (23), we can calculate the variance of the equivalent additive noise \( n_O (m) \):

\[\sigma_{NO}^2 = 2 \cdot \sum_{r=0}^{R-1} \int_{x=-\infty}^{\infty} (q_r - x)^2 \cdot \Pr (q_r | x) \cdot p_S (x) \cdot dx = \frac{\Delta^2}{3} \quad (37)\]

The resulting equivalent noise variance of the DAC with dither is 3 dB higher than that of a conventional DAC. This 3 dB degradation of the noise power is the price we pay in order to make the quantization noises across all the complex DACs within the array uncorrelated with each other. The benefit of this becomes apparent in the next section.
DACs (39) is upper bounded by \( k \) that of a signal arriving at user \( user \) where the channel is line-of-sight (LoS)

According to the statement proven in Appendix C, the ratio to \( nario \) when the transmitted signal of each antenna is equal case is single-user maximum ratio transmitter (MRT) sce-

The worst-case angles of departure are equal to:

\[
\theta_0 = 0 \quad \text{or} \quad \pm \pi/2 \quad \text{or} \quad \pm \pi/3 \quad (43)
\]

From (40), we see that for \( M > 2 \), the EVM of the MIMO array with dithered DACs is \( M/2 \) better than that of the conventional DACs’ upper bound (worst case), and the latter defines the necessary DAC resolution. To always satisfy the same EVM requirement, the array of \( M \) complex DACs with optimal dither can tolerate a quantization step that is \( \sqrt{M/2} \) as large as that of the conventional DACs, with the same size-\( M \) array.

Therefore, for a fixed EVM requirement, the array of \( M \) DACs with optimal dither needs \( \log_2(M/2) \) fewer bits than the same array with conventional DACs.

D. SIMULATION RESULTS

We illustrate the concept with the following numerical experiment.

- We simulate a downlink MIMO transmitter equipped with an \( M \) antenna uniform linear array, transmitting to \( K \) users.
- Each antenna is equipped with a pair of \( N \)-bit DACs.
- We assume that the signal \( x_k \) that we send to user \( k \) has a Gaussian distribution with identical variance \( \sigma^2 \). This assumption is typical for most popular OFDM modulation.
- The channel of each user is an LoS channel (42). The angle of departure of each user \( k \) is equal to:

\[
\alpha_k = (k \cdot \pi/K) + \theta_l, \quad k = 0, \ldots, K - 1 \quad (44)
\]

where the initial angular offset \( \theta_l \) is a simulation parameter.

\[
0 \leq \theta_l < \left( \frac{\pi}{K} \right) \quad (45)
\]

- We assume a zero-forcing transmitter. The transmitted signal vector with length \( M \) and elements \( s_l (m) \) is equal to,

\[
S_l = H^H \cdot \left( H^H \cdot H \right)^{-1} \cdot X \quad (46)
\]

where \( X \) is the users signal vector with length \( K \) and elements \( x_k \). \( H \) is the channel matrix with size \( K \times M \) and elements \( h_k (m) \). \( (.)^H \) and \( (.)^{-1} \) denote matrix complex conjugate transpose and matrix inversion operations, respectively.

- To avoid the clipping effect, we set the ratio between the squared DAC maximal value and input signal power to 15 dB.

\[
\text{Peak2rms} \triangleq \frac{2 \cdot q_0^2}{s_l^2 (m)} = \frac{M \cdot q_0^2}{K \cdot \sigma_X^2} = 15 \text{ dB} \quad (47)
\]

- \( K, N \) and \( M \) are simulation parameters.
- We measure the average EVM of the signal at each user.

Figure 4 presents the EVM of the signal arriving at the user versus the angle of departure for a single-user scenario.
and an array of 1000 antenna elements, each equipped with either 6-bit conventional DACs or 6-bit DACs with dither. For reference, we also present the analytical curve of the dithered DAC EVM calculated according to (39) as:

$$EVM_k = M \cdot \Delta^2 / \left(3 \cdot \sigma_X^2\right) = \text{Peak 2rms} / \left(3 \cdot 2^{2(N_{\text{bit}} - 1)}\right)$$  

(48)

Figure 5 presents the worst-case (single user and angle of departure is 0) EVM of a signal arriving at the user from a MIMO array with 1, 10, 100, 1000 and 10000 antenna elements, each equipped with a pair of dithered DACs, as a function of the DACs’ resolution. For reference we also present the worst-case performance of a convention DAC that is independent of the number of antenna elements.

From figures 4 and 5, we can make the following observations:

- The analytical EVM curve of the DAC with dither fully matches the simulation results.
- The performance of the DAC with dither does not depend on the user’s angles of departure; however, the performance of the conventional DAC does.
- The worst-case angles of departure satisfy (31).
- For a single-antenna transmitter, the DAC with dither causes a performance degradation of $10 \cdot \log_{10} \left(\frac{M}{2}\right)$ dB.
- For $M > 2$ number of antennas, in the worst-case scenario for the conventional DAC, the DAC with dither provides a gain equal to $10 \cdot \log_{10} \left(\frac{M}{2}\right)$ dB.
- For a single user, even in a typical (nonworst) case, the EVM performance of the DAC with dither is significantly better than the conventional one, given the same resolution.

From figure 6, we can conclude that for the multiuser case, the performance of the system employing dithered DACs decreases as the number of users increases. The amount of degradation for a $K$-user system is $10 \cdot \log_{10} K$ dB. This occurs because the power of each user decreases by $10 \cdot \log_{10} K$ dB, while the dither power stays constant. In contrast, the EVM performance of the conventional DAC improves with the increased number of users. It happens because the signal of each user serves as a dither to other users. We can see from figure 6, that at low resolution, the system employing the dithered DAC outperforms that of the conventional one. At high resolution, the conventional DAC can outperform the dithered one. However, a communication standard (e.g., 3GPP) typically mandates that the EVM not to exceed a certain limit across all usage scenarios. Therefore, DAC resolution should always be designed for the
worst-case single-user EVM, where DAC dithering significantly helps.

IV. UPLINK MIMO SYSTEM MODEL

Let us now consider an 'all digital' uplink MIMO receiver array equipped with $M$ antennas that serves $K$ users. Let us assume for simplicity a flat fading channel. We can always extend our results to a nonflat (frequency-selective) fading channel by adding the dimension of time to all our equations. For a flat fading channel, the output of ADC’s array is given by:

$$S_O = Q(S_I + N_I)$$  \hspace{1cm} (49)

Or in scalar form:

$$s_O (m) = Q(s_I (m) + n_I (m))$$  \hspace{1cm} (50)

where $S_O$, $S_I$ and $N_I$ are the ADC output signal, the ADC desired input signal and the input noise vectors with length $M$ and with elements $s_O (m)$, $s_I (m)$ and $n_I (m)$, respectively.

The desired signal vector is equal to:

$$S_I = H \cdot X$$  \hspace{1cm} (51)

Or in scalar form:

$$s_I (m) = \sum_{k=1}^{K} h_k (m) \cdot x_k$$  \hspace{1cm} (52)

where $X$ is the input symbol vector with length $K$ and elements $x_k$, and $H$ is the channel matrix with a size of $M \times K$ with elements $h_k (m)$.

We assume that the additive input noise $n_I (m)$ on each quantizer is an independent circularly symmetric complex Gaussian random variable with zero mean and identical variance $\sigma_N^2$. It is a valid assumption for the receiver thermal noise. The real and imaginary parts of the additive input noise $n_I (m)$ on each quantizer have an identical Gaussian distribution $p_{N} \sim N (0, \sigma_N^2 / 2)$.

We also assume that real and imaginary parts of each user signal $x_k$ are uncorrelated circularly symmetric complex random variables with zero mean and identical variance $\sigma_X^2$.

With these two assumptions in mind, we apply CEM to the uplink MIMO system based on (17):

$$S_O = F (S_I) + N_O$$  \hspace{1cm} (53)

Or in scalar form:

$$s_O (m) = F (s_I (m)) + n_O (m)$$  \hspace{1cm} (54)

where $N_O$ is the equivalent output noise vector with length $M$ and elements $n_O (m)$.

1) EQUIVALENT TRANSFER FUNCTION OF SINGLE ADC WITH GAUSSIAN INPUT NOISE

When input noise $n_I (m)$ has a Gaussian distribution, the quantizer equivalent transfer function and energy function are given by (7) and (16), where according to (8) and (3):

$$Pr (q_r | x) = 0.5 \times \begin{cases} 1 + \text{erf} \left( \frac{q_r + (\Delta / 2) - x}{\sigma_N} \right) & \text{if } (r = 1) \\ 1 - \text{erf} \left( \frac{q_r - (\Delta / 2) - x}{\sigma_N} \right) & \text{elseif } (r = R) \\ \text{erf} \left( \frac{q_r + (\Delta / 2) - x}{\sigma_N} \right) & \text{else} \end{cases}$$  \hspace{1cm} (55)

where $\text{erf} (\cdot)$ denotes the Gaussian error function.

Therefore, the equivalent transfer and energy functions are equal to:

$$F_R (x) = q_1 \cdot (1 + \text{erf} \left( \frac{(q_1 + (\Delta / 2) - x)}{\sigma_N} \right)) + q_1 \cdot (1 + \text{erf} \left( \frac{(q_1 + (\Delta / 2) - x)}{\sigma_N} \right)) + \sum_{r=2}^{R-1} q_r \cdot \left( \text{erf} \left( \frac{q_r + (\Delta / 2) - x}{\sigma_N} \right) - \text{erf} \left( \frac{q_r - (\Delta / 2) - x}{\sigma_N} \right) \right)$$  \hspace{1cm} (56)

$$V_R (x) = q_1^2 \cdot (1 + \text{erf} \left( \frac{(q_1 + (\Delta / 2) - x)}{\sigma_N} \right)) + q_1^2 \cdot (1 + \text{erf} \left( \frac{(q_1 + (\Delta / 2) - x)}{\sigma_N} \right)) + \sum_{r=2}^{R-1} q_r^2 \cdot \left( \text{erf} \left( \frac{q_r + (\Delta / 2) - x}{\sigma_N} \right) - \text{erf} \left( \frac{q_r - (\Delta / 2) - x}{\sigma_N} \right) \right)$$  \hspace{1cm} (57)

Figures 7 and 8 present the equivalent transfer function $F_R (s_I)$ (7) for 1- and 2-bit ADCs (on either the real or imaginary part of the input), respectively, at different normalized noise variances $(2 \cdot \sigma_N / \Delta)$. It will be seen that additive noise with enough variance has a linearizing effect on the ADC equivalent transfer function.

V. UPLINK MIMO RECEIVER ADC RESOLUTION DETERMINATION

To determine the resolution of an ADC array serving multiple uplink users, we assume that the desired input signal $s_I (m)$ on each antenna is a circularly symmetric complex Gaussian random variable with zero mean and variance

$$\sigma_S (m)^2 = \frac{1}{2} \cdot E \left[ s_I (m)^2 \right] = \frac{1}{2} \cdot \sum_{k=1}^{K} |h_k (m)|^2 \cdot \sigma_X^2$$  \hspace{1cm} (58)

This assumption is correct when user signals $x_k, k = 0, \ldots, K - 1$ are complex random variables with an identical Gaussian distribution, which is true for most popular OFDM modulations. For non-OFDM signals, according to the central
limit theorem, when the number of users $K$ is sufficiently large, the PDF of an antenna’s desired signal $s_l (m)$ is also approximately Gaussian.

Let us further assume for simplicity that the variance of the desired signal $s_l (m)$ is constant for all antennas

$$
\sigma_S (m)^2 = \sigma_S^2, \quad \text{for any } m.
$$

(59)

This is true for the line-of-sight (LoS) channel, which is typical for mmWave systems that represent one of the most important practical applications of massive MIMO. This assumption is also valid when the signal bandwidth is sufficiently large to absorb all frequency-selective fading energy fluctuations.

From assumption (59), it follows that the real and imaginary parts of the desired input signal $s_l (m)$ on each antenna’s quantizer are uncorrelated and have an identical Gaussian distribution $p_S \sim N \left(0, \sigma_S^2 / 2 \right)$.

### A. NONLINEAR DISTORTION

Assuming the desired signal and the input noise both have a Gaussian distribution, we can apply the Bussgang-Rowe decomposition [5], [6] to represent the MIMO receiver input vector $S^r$ as a sum of the desired signal vector $S_l$ scaled by a certain gain $g$, equivalent additive white noise vector $N_o$ and NLD vector $W_o$.

$$S^r = Q (S_l + N_l) = F (S_l) + N_o = g \cdot S_l + W_o + N_o$$

(60)

where NLD vector has length $M$ and elements $w_o (m)$ defined as:

$$W_o \triangleq F (S_l) - g \cdot S_l$$

(61)

According to the Bussgang theorem [5], [6] the scalar gain is equal to:

$$g = \frac{2}{\sigma_S^2} \cdot \int_{x=-\infty}^{+\infty} x \cdot F_R (x) \cdot p_S (x) \cdot dx$$

(62)

The NLD vector has no correlation with the desired signal:

$$E \left[ S_l \cdot W_o^H \right] = 0$$

(63)

From the NLD definition (61) and properties (12) and (13) it also follows:

$$E \left[ N_o \cdot W_o^H \right] = 0$$

(64)

According (61) and (63), the diagonal elements of the NLD autocorrelation matrix $E \left[ W_o \cdot W_o^H \right]$ are equal to,

$$\sigma_{W_o}^2 = E \left[ |w_o (m)|^2 \right] = E \left[ |F (s_l (m))|^2 \right] - g_o^2 \cdot E \left[ |s_l (m)|^2 \right]$$

$$= 2 \cdot \int_{x=-\infty}^{+\infty} F_R (x)^2 \cdot p_S (x) \cdot dx - g_o^2 \cdot \sigma_S^2$$

(65)

However, we cannot assume that the NLD is necessarily white:

$$E \left[ W_o \cdot W_o^H \right] \neq \sigma_{W_o}^2 \cdot I$$

(66)

**To Summarize:** Our new equivalent model says the output of the quantizer can be represented as the sum of a desired signal multiplied by an equivalent gain, with additive white noise and nonwhite NLD.

The separation of quantizer output deviation from the desired signal on the white and nonwhite components is important because as we will show in the sequel, these two components affect the MIMO system differently.

### B. CONNECTION BETWEEN THE CEM-BASED NLD MODEL AND CONVENTIONAL BUSSGANG-ROWE MODEL

Note that we apply the Bussgang-Rowe decomposition to the desired signal only; in contrast, the widely used approach of ADC modeling (see [28]–[30] and [51]) currently applies the Bussgang-Rowe decomposition [5], [6] to the sum of the

---

**FIGURE 7.** The equivalent transfer function of the 1-bit ADC.

**FIGURE 8.** The equivalent transfer function of the 2-bit ADC.
desired signal and the input noise in order to represent the quantizer output as:

\[ S_O = Q (S_I + N_I) = g_B \cdot (S_I + N_I) + W_B \]  

(67)

where the gain and NLD according to this decomposition are equal to:

\[ g_B = \frac{2}{\sigma_S^2} \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y) \cdot Q_R (x + y) \cdot p_S (x) \cdot p_N (y) \cdot dy \cdot dx \]  

(68)

\[ W_B \equiv S_O - g_B \cdot (S_I + N_I) \]  

(69)

We refer to (67) as the conventional Bussgang decomposition. The corresponding NLD vector \( W_B \) also has properties:

\[ E \left[ N_I \cdot W_B^H \right] = E \left[ S_I \cdot W_B^H \right] = 0 \]  

(70)

According to (69), the diagonal elements of the Bussgang NLD autocorrelation matrix \( E \left[ W_B \cdot W_B^H \right] \) are equal to,

\[ \sigma_{W_B}^2 = E \left[ |w_B(m)|^2 \right] = E \left[ Q (s_I(m) + n_I(m)) |^2 \right] - g_B^2 \cdot E \left[ |s_I(m) + n_I(m)|^2 \right] 
= 2 \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} Q_R (x + y)^2 \cdot p_S (x) \cdot p_N (y) \cdot dy \cdot dx 
- g_B^2 \cdot \left( \sigma_S^2 + \sigma_N^2 \right) \]  

(71)

However, we cannot assume that the NLD is Bussgang necessarily white:

\[ E \left[ W_B \cdot W_B^H \right] \neq \sigma_{W_B}^2 \cdot I \]  

(72)

From (62), (68) and (7), it follows that the CEM equivalent gain \( g_O \) and Bussgang gain \( g_B \) are identical: \( g_O = g_B \). Then:

\[ S_O = g_O \cdot S_I + N_O + W_O = g_O \cdot (S_I + N_I) + W_B \]  

(73)

Comparing (69) and (61), our NLD turns out to be the conditional expectation of the conventional Bussgang NLD

\[ W_O = F (S_I) - g_O \cdot S_I = E \left[ W_B | S_I \right] \]  

(74)

It means that the conventional Bussgang NLD has a nondeterministic, white component:

\[ N_B \triangleq W_B - E \left[ W_B | S_I \right] = W_B - W_O \]  

(75)

It has this white component because the conventional Bussgang NLD is a function of the sum of the desired signal and input white noise.

We can consider the CEM output white noise to consist of the sum of the scaled input noise and the white noise contained in the conventional Bussgang NLD, as follows:

\[ N_O = g_O \cdot N_I + N_B, \quad W_O = W_B - N_B \]  

(76)

Or in scalar form:

\[ E \left[ |n_O (m)|^2 \right] = g_B^2 \cdot E \left[ |n_I (m)|^2 \right] + E \left[ |n_B (m)|^2 \right] \]  

(77)

\[ E \left[ |w_O (m)|^2 \right] = E \left[ |w_B (m)|^2 \right] - E \left[ |n_B (m)|^2 \right] \]  

(78)

It means that new model more accurately separates white and nonwhite components of the quantizer output deviation from the desired signal. It is important because as we will show in the sequel, these two components affect MIMO system differently.

Another advantage of CEM-based Bussgang decomposition over the conventional Bussgang decomposition is that in contrast to the common approach for which NLD \( W_B \) is the function (69) of the sum \( (S_I + N_I) \), the new equivalent model presents NLD as the function (61) of the desired signal \( S_I \) only. It is much more convenient for NLD suppression. The general block diagram of the iterative receiver with CEM-based NLD cancellation is shown in figure 9, where \( \hat{X} \), \( \hat{S}_I \) and \( \hat{W}_O \) are estimation of data, desired signal and NLD vectors, respectively, and ‘naïve’ receiver means that the receiver does not consider the effects of low-resolution ADC.

![FIGURE 9. The CEM-based NLD suppression receiver block diagram.](image)

An example of such a receiver will be given in Section VI.

We can also implement the NLD suppression receiver based on the conventional Bussgang-Rowe decomposition. However, because the NLD is a function of the sum of the desired signal and noise in conventional Bussgang-Rowe decomposition, we have to cancel from the ADC output signal the expectation of NLD conditioned on the predicted desired signal, as shown in the figure below.

This approach has high computational complexity since it requires calculation of the conditional expectation of NLD for each sample and each iteration. We can reduce this complexity by implementing the expectation function as a precalculated lookup table. According to the CEM model (6), this lookup table is the equivalent transfer function of the ADC.

### C. ONE-BIT ADC CASE

From (56) and (57), it follows that for the special case of a 1-bit ADC, the equivalent transfer and energy function are equal to,

\[ F_R (x) = \Delta / 2 \cdot \text{erf} \left( x / \sigma_N \right) \quad \text{and} \quad V_R (x) = \left( \Delta / 2 \right)^2 \]  

(79)

According to the error function integral table given in [7] and from (62) and (68), we can express the 1-bit ADC gain as:

\[ g_O = g_B = \frac{2}{\sigma_S^2} \cdot \int_{-\infty}^{+\infty} x \cdot F_R (x) \cdot p_S (x) \cdot dx \]

\[ = \frac{\Delta}{2} \cdot \sqrt{\frac{2 / \pi}{\sigma_S^2 + \sigma_N^2}} \]  

(80)
where the distribution of the desired signal is normal:

\[ p_S(x) = \left( \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_S^2}} \right) \cdot \exp \left( -x^2 / \left( 2 \cdot \sigma_S^2 \right) \right) \]  

From (15) and (65), we can express the variance of the ADC output equivalent white noise and NLD according to the CEM-based Bussgang decomposition,

\[ \sigma_{NO}^2 = E \left[ |n_O(m)|^2 \right] = \frac{\Delta^2}{2} - 2 \cdot \int_{-\infty}^{+\infty} F_R(x)^2 \cdot p_S(x) \cdot dx \]

\[ = \left( \frac{\Delta^2}{2} \right) \cdot \left( 2 - \left( \frac{4}{\pi} \right) \cdot \tan^{-1} \left( \frac{1}{\sqrt{1 + 2 \cdot \left( \frac{\sigma_S^2}{\sigma_N^2} \right)}} \right) \right) \]  

\[ \sigma_{WO}^2 = E \left[ |w_O(m)|^2 \right] = 2 \cdot \int_{-\infty}^{+\infty} F_R(x)^2 \cdot p_S(x) \cdot dx - \sigma_S^2 \cdot \sigma_N^2 \]

\[ = \left( \frac{\Delta^2}{2} \right) \cdot \left( \frac{4}{\pi} \right) \cdot \tan^{-1} \left( \frac{1}{\sqrt{1 + 2 \cdot \left( \frac{\sigma_S^2}{\sigma_N^2} \right)}} \right) - 1 \]

\[ - \sigma_S^2 \cdot \sigma_N^2 \]  

From (87), we can express the variance of the ADC output white noise and NLD according to the conventional Bussgang decomposition,

\[ E \left[ |g_O \cdot n_I(m)|^2 \right] = g_O^2 \cdot \sigma_N^2 \]

\[ E \left[ |w_B(m)|^2 \right] = \left( \frac{\Delta^2}{2} \right) - g_O^2 \cdot \left( \sigma_S^2 + \sigma_N^2 \right) \]  

From (80), (82), (83), (84) and (85), we derive the variance of the white component that CEM-based Bussgang decomposition extracts from the conventional Bussgang decomposition NLD,

\[ \sigma_{\Delta}^2 = E \left[ |w_B(m)|^2 \right] - E \left[ |w_O(m)|^2 \right] \]

\[ = \left( \frac{\Delta^2}{2} \right) - \left[ \sigma_S^2 \cdot \sigma_N^2 \right] \]

\[ = \left( \frac{\Delta^2}{2} \right) \cdot \left( 2 - \frac{4}{\pi} \cdot \tan^{-1} \left( \frac{1}{\sqrt{1 + 2 \cdot \left( \frac{\sigma_S^2}{\sigma_N^2} \right)}} \right) - 2 \cdot \frac{\sigma_N^2}{\sigma_S^2 + \sigma_N^2} \right) \]  

It means that when the input SNR approaches zero, the conventional Bussgang decomposition NLD contains only the white component:

\[ N_O = g_O \cdot N_I + W_B, \quad W_O = 0 \]  

It confirms the well-known fact that when the input SNR approaches zero, the conventional Bussgang decomposition NLD autocorrelation matrix is diagonal [28] and equal to:

\[ E \left[ W_B \cdot W_B^H \right] = \left( \Delta/2 \right) \cdot \left( 1 - \left( \frac{2}{\pi} \right) \right) \cdot I \]  

D. SINGLE ADC PERFORMANCE

Let us define the ADC input signal to noise ratio (SNR) as the ratio between the desired signal and noise variances and the real ADC scaling factor (SF) as the ratio between the ADC saturation level and the input signal variance,

\[ SNR_{ADC,I} \triangleq \frac{\sigma_S^2}{\sigma_N^2} \]

\[ SF \triangleq \frac{2}{\sqrt{\pi}} \frac{\sigma_N^2}{\sigma_S^2} \]  

According to assumption (59), each antenna’s ADC input SNR and SF are identical.

According to CEM, the ADC output can be represented as the sum of a desired signal multiplied by an equivalent gain, with equivalent additive white noise and nonwhite NLD. Let us define the ADC output SNR, SDR (signal to distortion ratio) and SINAD (signal to noise and distortion ratio) as:

\[ SNR_{ADC,O} \triangleq \frac{\sigma_S^2 \cdot \sigma_N^2}{\sigma_N^2} \]

\[ SDR_{ADC,O} \triangleq \frac{\sigma_S^2 \cdot \sigma_N^2}{\sigma_WO^2} \]

\[ SINAD_{ADC,O} \triangleq \frac{\sigma_S^2 \cdot \sigma_N^2}{\sigma_WO^2} \]  

where the equivalent gain \( g_O \), equivalent additive white noise variance \( \sigma_{NO}^2 \) and NLD variance \( \sigma_{WO}^2 \) can be calculated according to (62), (65), (15), (6) and (16). These parameters are a function of \( p_S \sim N \left( 0, \frac{\sigma_S^2}{2} \right) \) and \( n_I \sim N \left( 0, \frac{\sigma_N^2}{2} \right) \). From (92), we can express the desired signal and input noise variance as a function of the ADC input SNR and scaling factor:

\[ \sigma_{\Delta}^2 = 2 \cdot q_0^2 \left( \frac{SF \cdot \left( 1 + SNR_{ADC,I} \right)}{\sigma_N^2} \right) \]

\[ \sigma_{\Delta}^2 = SNR_{ADC,I} \cdot \sigma_N^2 \]  

Therefore, we can express the ADC output SNR, SDR and SINAD as functions of the ADC input SNR and scaling factor.

Figures 11, 12 and 13 show a single ADC output SNR, SDR and SINAD as functions of the input SNR calculated for the optimal SF that maximizes the ADC output SINAD.

\[ SF_{OPTIM} = \arg \max_{SF} \left( SINAD_{ADC,O} \left( SF, SNR_{ADC,I} \right) \right) \]  

To obtain this curve for a 1-bit ADC, we calculate \( g_O, \sigma_{NO}^2, \sigma_{WO}^2 \) from the analytical expressions (80), (82) and (83), respectively. For an ADC with larger than one-bit resolution,
to calculate integrals (62), (65) and (15), we apply the numerical integration approximation:

\[
p_s(s) \approx \sum_{n=-N}^{+N} \Delta s \cdot p_S(n \cdot \Delta s) \cdot \delta(s-n \cdot \Delta s) = N \cdot \frac{1}{\Delta s} \cdot \frac{\int p_S(s) ds}{\Delta s}
\]  

where \( \Delta s = 0.01 \cdot \min(\sigma_S, \sigma_N) \), \( N = \text{ceil} \left( \frac{5 \cdot \sqrt{\sigma_S^2 + \sigma_N^2}}{\Delta s} \right) \) and \( \delta(s) \) denotes the Dirac function.

The optimal SF was obtained by searching over all SF from 0 to 30 dB in 0.1 dB steps.

For reference, for 2-, 3-, 4- and 5-bit ADC, we also provide the SINAD curve calculated according to the conventional model assuming additive quantization noise variance of \( \Delta^2 / 12 \)

\[
\text{SINAD}_{ADC,CONV} \triangleq \sigma_S^2 / \left( \sigma_N^2 + \left( \Delta^2 / 12 \right) \right)
\]  

(101)

From these figures, we can conclude that the ADC output SDR is a monotonically increasing function of the input SNR that converges to a certain fixed value when the input SNR approaches infinity. In contrast, the ADC output SNR is a monotonically decreasing function of the input SNR.

In the low-input SNR regime, the resulting SINAD is dictated by the output white noise, and in the high-input SNR regime, the resulting SINAD is dictated by the output NLD.

The higher the ADC resolution is, the closer the resulting SINAD calculated by the CEM model (96) to the resulting SINAD calculated by the conventional model. However, the main advantage of the CEM model over the conventional model is that the CEM model provides separation between the white and nonwhite part of the output noise.

### E. ADC ARRAY NOISE FIGURE

Now, we apply the single ADC performance result to that of an array of ADCs and find the maximal performance degradation of a MIMO receiver caused by the finite resolution of the ADCs making up the array. We refer to this quantity of the ADC array as the worst-case noise figure (NF) in the sequel. We will show that the worst-case NF is a decreasing function of ADC resolution. Our goal is to determine the minimal ADC resolution required to keep the ADC array’s worst-case NF at an acceptably low level.

Common uplink receiver options for detecting the user signal vector \( X \) without concern for quantization error are listed in Table 1, where \((\cdot)^H\) and \((\cdot)^{-1}\) denote matrix conjugate transpose and matrix inversion operations, respectively, \( \text{diag} (\cdot) \) denotes selecting the matrix diagonal elements only,
TABLE 1. Different receiver options.

| Receiver Type | Operation |
|---------------|-----------|
| MRC           | \( \hat{X} = \text{diag} \ A^{-1} \cdot Y \) |
| ML            | \( \hat{X} = \arg \min_x \ Y - A' \cdot x \cdot A \cdot x - A' \cdot y \) |
| ZF            | \( \hat{X} = A' \cdot y \) |
| MMSE          | \( \hat{X} = A' \cdot y \) |

\( I \) is the identity matrix, \( A \) is the autocorrelation matrix and \( Y \) is the maximum ratio combining (MRC) vector.

\[
A = H^H \cdot H \quad (102) \\
Y = H^H \cdot S_O \quad (103)
\]

The table includes the maximum ratio combining (MRC), maximum likelihood (ML), zero forcing (ZF) and minimum mean square error (MMSE) receivers.

The MRC vector \( Y \) is the common front-end for the other receiver types. We can always model the MRC output variance increase as an equivalent increase of the input additive white noise. Therefore, the NF of the MRC receiver upper bounds the NF of the other receiver types previously mentioned. Consequently, if the ADC resolution is sufficient for the MRC receiver, it will also be sufficient for the other receiver types. For this reason, our paper only evaluates the NF of the MRC receiver.

The output of the MRC receiver with an ideal ADC without quantization error is,

\[
y_{k,I} = \tilde{y}_k + \sum_{m=1}^{M} h_k (m)^* \cdot n_l (m) \quad (104)
\]

where \( \tilde{y}_k \) is the desired output of the MRC receiver for user \( k \).

\[
\tilde{y}_k = \sum_{m=1}^{M} h_k (m)^* \cdot s_I (m) \quad (105)
\]

The receiver with finite-resolution ADCs must normalize the MRC result by the equivalent ADC gain \( g_O \).

\[
y_{k,Q} = (1/g_O) \cdot \sum_{m=1}^{M} h_k (m)^* \cdot s_O (m) \quad (106)
\]

Let us define the ADC array noise figure (NF) as the ratio between the variances of the deviation of the MRC output from the desired value of MIMO receivers equipped with finite-resolution ADCs and with ideal (infinite resolution with no quantization error) ADCs.

\[
NF (k) \equiv \frac{E \left[ |y_{k,Q} - \tilde{y}_k|^2 \right]}{E \left[ |y_{k,I} - \tilde{y}_k|^2 \right]} \quad (107)
\]

Let us define the worst-case NF as the maximal NF over all users and all possible channel realizations:

\[
NF_{\text{MAX}} \equiv \max_{H,k} (NF (k)) \quad (108)
\]

According to (104), for ideal MRC receivers:

\[
E \left[ |y_{k,I} - \tilde{y}_k|^2 \right] = \sum_{m=1}^{M} |h_k (m)|^2 \cdot \sigma_N^2 \quad (109)
\]

According to the CEM formulation, for finite-resolution MRC receivers

\[
E \left[ |y_{k,Q} - \tilde{y}_k|^2 \right] = E \left[ |n_{O,k}|^2 \right] + E \left[ |w_{O,k}|^2 \right] \quad (110)
\]

where \( n_{O,k} \) and \( w_{O,k} \) are post-MRC additive white noise and the NLD of the \( k \)-th user, respectively.

\[
n_{O,k} = (1/g_O) \cdot \sum_{m=1}^{M} h_k (m)^* \cdot n_O (m) \quad (111)
\]

\[
w_{O,k} = (1/g_O) \cdot \sum_{m=1}^{M} h_k (m)^* \cdot w_O (m) \quad (112)
\]

Since, according to property (14), the equivalent output noise \( n_{O,k} (m) \) is a white process, the corresponding post-MRC equivalent noise variance is equal to:

\[
\sigma^2_{O,k} = E \left[ |n_{O,k}|^2 \right] = \frac{1}{g_O} \cdot \sum_{m=1}^{M} |h_k (m)|^2 \cdot \sigma^2_{NO} \quad (113)
\]

However, the NLD is not necessarily a white process and in the worst case, may sum coherently.

According to (111) and the property proven in Appendix B, the worst-case post-MRC NLD variance is equal to:

\[
\sigma^2_{NO,k} = \max (\sigma^2_{W_O,k}) = \frac{\left( \sum_{m=1}^{M} |h_k (m)| \right)^2}{g^2_O} \cdot \sigma^2_{W_O} \quad (114)
\]

According to (109), (113) and (114), the worst-case ADC array noise figure (NF) as defined above for a user \( k \) is now:

\[
NF (k) = \frac{\left( \sum_{m=1}^{M} |h_k (m)| \right)^2}{g^2_O \cdot \sigma^2_N} + \left( \frac{\left( \sum_{m=1}^{M} |h_k (m)| \right)^2}{g^2_O \cdot \sigma^2_{NO}} \right) \cdot \sigma^2_{W_O} \quad (115)
\]

According to the property proven in Appendix C:

\[
NF_{\text{MAX}} = \max (NF (k)) = \frac{\left( \sigma^2_{NO} + M \cdot \sigma^2_{W_O} \right)}{\left( g^2_O \cdot \sigma^2_N \right)} \quad (116)
\]

As we can see from this expression, the worst-case NF is independent of the user index \( k \) and always equal to a constant value \( NF_{\text{MAX}} \).

In Appendix D, we provide an example of such a worst-case scenario consisting of a single user with an LoS channel:

\[
h_0 (m) = g_0 \cdot \exp (j \cdot \pi \cdot \sin (\alpha_0) \cdot m) \quad (117)
\]

The worst-case angle of arrival is equal to:

\[
\alpha_0 = 0 \quad \text{or} \quad \pm \pi/2 \quad \text{or} \quad \pm \pi/3 \quad (118)
\]
The equivalent gain $g_O$, the equivalent additive white noise variance $\sigma^2_{NO}$ and NLD variance $\sigma^2_{WO}$ can be calculated according to (62), (65), (15), (6) and (16). These parameters are functions of $p_S \sim N \left(0, \frac{\sigma^2_S}{2}\right)$ and $p_X \sim N \left(0, \frac{\sigma^2_N}{2}\right)$.

Let us define the $k$-th user input SNR as the post-MRC SNR of the ideal receiver,

$$SNR_{UE,I} (k) = \sum_{m=1}^{M} |h_k (m)|^2 \cdot \frac{\sigma^2_X}{\sigma^2_N}$$  (119)

Let us also define the cumulative input SNR as the sum of the input SNR of all users, which according to (59) and (58), is equal to

$$SNR_{\sum,I} \triangleq \sum_{k=1}^{K} SNR_{UE,I} (k) = \sum_{k=1}^{K} \sum_{m=1}^{M} |h_k (m)|^2 \cdot \frac{\sigma^2_X}{\sigma^2_N} = M \cdot \frac{\sigma^2_S}{\sigma^2_N}$$  (120)

From (120), we may express the desired signal and input noise variance as a function of the cumulative input SNR and scaling factor that according to (59), is common for all ADCs:

$$\sigma^2_N = 2 \cdot q^2_0 \left( (SF \cdot (1 + (SNR_{\sum,I}/M))) \right)$$  (121)

$$\sigma^2_S = (SNR_{\sum,I}/M) \cdot \sigma^2_N$$  (122)

Therefore, we can express the ADC array worst-case noise figure $N_{MAX}$ as a function of the ADC cumulative input SNR and scaling factor.

From (80), (82) and (83), it follows that for a 1-bit ADC:

$$\lim_{\sigma^2_N \rightarrow 0} (g_O) = \frac{\Delta}{\sqrt{\pi \cdot \sigma^2_N}}, \quad \lim_{\sigma^2_N \rightarrow 0} \left( \sigma^2_{WO} \right) = 0,$$

$$\lim_{\sigma^2_N \rightarrow 0} \left( \sigma^2_{NO} \right) = \frac{\Delta^2}{2}$$  (123)

Therefore, from (115), it follows that:

$$\lim_{SNR_{\sum,I} \rightarrow 0} \left( N_{MAX} \left( SNR_{\sum,I} \right) \right) = \lim_{\sigma^2_N \rightarrow 0} \left( \frac{\sigma^2_{NO} + M \cdot \sigma^2_{WO}}{g^2_0 \cdot \sigma^2_N} \right) = \frac{\pi}{2} \approx 1.96dB$$  (124)

which confirms the well-known conclusion ([26] and [51]) that in the low SNR regime, the performance degradation caused by a 1-bit ADC converges to 1.96 dB.

To obtain the curve of the ADC array worst-case noise figure for a 1-bit ADC, we calculate $g_O$, $\sigma^2_{NO}$, and $\sigma^2_{WO}$ from analytical expressions (80), (82) and (83), respectively. For an ADC with a higher than one-bit resolution, to obtain these curves, we apply numerical integration approximation (100). The ADC array worst-case noise figure is calculated for optimal SF.

$$SF_{OPTIM} = \arg \max_{SF} \left( N_{MAX} \left( SF, SNR_{\sum,I} \right) \right)$$  (125)

The optimal SF was obtained by searching over all SF from 0 to 30 dB with a step of 0.1 dB.

For reference, we also provide the UE output signal to noise and distortion ratio (SINAD) as a function of the UE input SNR for the scenario when we have single-user, LoS channel with an angle of arrival equal to (118) since it is the worst case for the ADC array NF. For the single user, the cumulative input SNR is equal to the UE input SNR. The UE output SINAD is given by the expression:

$$\min \left( SNR_{UE,Q} (k) \right) = SNR_{UE,I} (k)/N_{MAX}$$  (126)

Figures 14, 15, 16 and 17 summarize various aspects of the results of this evaluation for ADC resolutions from 1 to 5 bits and for numbers of antennas employed by the ADC array at 1, 10, 100, 1000 and 10000.

We can see from these figures that when the cumulative input SNR is low, the increase in the cumulative input SNR causes a linear increase in the output SINAD, and the NF stays at a fixed low level. For a 1-bit ADC, this level is 1.96 dB, as was predicted in (124). However, as the cumulative input SNR grows beyond a certain limit, any additional increase in the cumulative input SNR starts to decrease the output SINAD, causing dramatic growth of NF. The reason for this is that when the SNR is low, the NLD is mild, and when the SNR is high, the NLD becomes dominant, as we have seen in Figure 9. The higher the ADC resolution and the greater the number of antennas are, the higher the threshold for the cumulative input SNR, below which a conventional MRC can operate without significant NF degradation.

Similar curves of single-user post-MRC output SINAD were presented in [8]. However, the results therein were obtained from numerical simulations, whereas our results are from numerical integration of analytical expressions. The two results match quite well.

**F. ADC RESOLUTION DETERMINATION METHODOLOGY**

Our equivalent model provides a practical design rule to determine the ADC resolution. Suppose we allow a 3-dB performance degradation when an infinite-resolution

$$SNR_{UE,Q} (k) = SNR_{UE,I} (k)/N_{MAX}$$  (126)
ADC array is replaced by a finite-resolution one, which means the maximal NF equals 3 dB, we can then compute the maximal allowable value of the cumulative input SNR accordingly.

\[
\text{SNR}_{\Sigma, \text{In}, 3dB} \triangleq \text{Solve}_{\text{SNR}_{\Sigma, \text{In}}} \left( \min \left( \text{NF}_{MAX} \left( \text{SF}, \text{SNR}_{\Sigma, \text{In}} \right) \right) = 3dB \right)
\]

(127)

where Solve \((f(x) = A)\) means finding the value of \(x\) that fulfills equality \(f(x) = A\) and NF given by equation (127).

We apply the numerical integration approximation (100) to evaluate the above cumulative input SNR 3-dB threshold according to (127) for a Gaussian input signal distribution, for different ADC resolutions and different numbers of antennas. The results are shown in figure 18.

As the figure illustrates, for a given number of antennas, a desired cumulative SNR requires a certain number of ADC bits.

Our methodology uses tables of the type illustrated by figure 18. First, we set the number of antennas and the maximal possible cumulative SNR. Second, we use Figure 16 to find the curve that lies just above the intersection point of the maximal possible cumulative SNR on the y-axis and the number of antennas on the x-axis. The number of bits associated with the curve is the required ADC resolution.

As an example, our methodology, assuming we have 10,000 receive antenna elements and a maximal possible cumulative SNR of 40 dB, indicates that according to figure 18, a 3-bit ADC is sufficient.
G. SIMULATION RESULTS
To confirm the theoretical analysis, we simulate a MIMO uplink receiver system with a single user and a base station equipped with 100 antennas. The modulation is OFDM with QAM64 on each subcarrier. MRC combining is employed. We simulate ADC with 1, 2, 3, 4 and 5 bits of resolution. The ADC input SF was optimized according to (125). The channel is LoS defined by (117). Figures 19 and 20 present bit error rate (BER) simulation results for a worst-case channel when the angle of arrival is 0° and for an average channel when the angle of arrival is a random variable uniformly distributed from 0 to π.

![Graph](image)

**FIGURE 19.** The BER for the worst-case channel.

![Graph](image)

**FIGURE 20.** The BER for an average channel.

We see that when the cumulative input SNR (post-MRC input SNR) is low and the nonlinear distortion is mild, increasing the input SNR improves the performance. However, after the cumulative input SNR exceeds a certain threshold, nonlinear distortion becomes dominant and further increases in the cumulative input SNR cause performance degradation. The higher the ADC resolution is, the greater the value of this threshold is. For the worst-case scenario, the value of this threshold (at least for the 1-bit, 2-bit, and 3-bit cases) matches our prediction in figure 16. The performance on average channels is somewhat better, but the same threshold applies.

H. FUTURE WORK
For ADC resolution determination, we assume a uniform distribution of desired signal energy between antennas (59). Future work aims to extend the analysis to a narrowband multipath scenario where a nonuniform distribution of the desired signal energy between antennas may cause an increase of NF. To take this problem into account, we propose to define the worst-case effective antenna number as

\[ \hat{M} = \frac{\sum_{m=1}^{M} \sigma_s(m)^2}{\max(\sigma_s(m)^2)} \]

We can use the effective antenna number instead of the actual antenna number to determine the ADC resolution. We define the channel efficiency coefficient as: \( \alpha = \hat{M} / M \). Then, the effective antenna number is equal to \( \hat{M} = \alpha \cdot M \). The channel efficiency coefficient \( \alpha \) is determined by channel measurements and should be part of the channel model.

VI. UPLINK MIMO RECEIVER CONSIDERING THE EFFECT OF LOW-RESOLUTION ADC
In Section V, it was shown that in the low SNR regime, NLD is negligible and even conventional MIMO receivers may work. However, beyond this range, NLD may cause significant performance degradation. We can extend the SNR working range by suppressing NLD. We view all MIMO receivers that consider the low-resolution ADC effect [31]–[43] as a certain type of NLD suppression receiver.

Several existing low-resolution ADC receiver algorithms, e.g., the successive-cancellation MIMO receiver presented in [40], employ a conditional expectation notion similar to the CEM in part of their derivation, though without providing formal justification.

The CEM model application significantly reduces the computational complexity of the MIMO receiver considering the effect of low-resolution ADC.

- Subsection A presents the CEM-based maximum likelihood ML MIMO receiver.
- Subsection B presents the CEM-based iterative MMSE MIMO receiver.

A. ML RECEIVER
1) EXISTING ML RECEIVER
We assume that each user signal \( x_k \) is an independent QAM signal with \( N_{QAM} \) possible values (constellation points).

If the ADC resolution is sufficiently high, we can neglect the quantization error and use a conventional (naïve) maximum likelihood (ML) MIMO receiver given by:

\[
\hat{X} = \arg\min_{\hat{X}} \left( (S_O - H \cdot \hat{X})^H \cdot (S_O - H \cdot \hat{X}) \right)
\]

\[
= \arg\min_{\hat{X}} \left( (Y - \hat{Y}(\hat{X}))^H \cdot R \cdot (Y - \hat{Y}(\hat{X})) \right) \tag{128}
\]
where \( Y = H^H \cdot S_O \) is the maximum ratio combining (MRC) output vector, \( \hat{Y} (\hat{X}) = H^H \cdot H \cdot \hat{X} \) is the MRC constellation lookup table precalculated for all \( N_{QAM}^K \) possibilities of vector \( \hat{X} \), and \( R = (H^H \cdot H)^{-1} \) is the precalculated inverse autocorrelation matrix of the channel.

The number of complex multiplications required by the naïve ML receiver is equal to:

\[
C = C_{MRC} + C_{DIST} \cdot N_{QAM}^K
\]

(129)

where \( C_{MRC} = M \cdot K \) is the MRC calculation complexity, and \( C_{DIST} = K^2 + K \) is the distance calculation complexity.

As presented in [35], the ML receiver that takes into account the quantization error chooses such a data vector \( X \) that maximizes the likelihood of the observed ADC output:

\[
\hat{X} = \arg \max_{\tilde{X}} \left( \Pr (S_O | \tilde{s}_I = H \cdot \tilde{X}) \right)
\]

= \arg \max_{\tilde{X}} \left( \prod_{m=1}^{M} \Pr (s_O (m) | s_I (m)) \right) \tag{130}

where:

\[
\Pr (s_O (m) | s_I (m)) = \Pr (\text{Re} (s_O (m)) | \text{Re} (s_I (m))) \times \Pr (\text{Im} (s_O (m)) | \text{Im} (s_I (m)))
\]

(131)

for the Gaussian noise conditional probability of ADC output, \( \Pr (q_l | x) \) is given by (55). Even if we assume that the error function calculation is implemented as a lookup table and we can neglect its complexity, the residual number of complex multiplications of the ML algorithm (130) is equal to:

\[
C \geq M \cdot C_{Pr} \cdot N_{QAM}^K
\]

(132)

where \( C_{Pr} = K + 1 \) is the complexity to calculate a single ADC output probability.

This is almost \( (M/K) \) times higher than (129), which makes it impractical.

2) CEM-BASED ML RECEIVER

According to the new equivalent model, the output of the MRC receiver is equal to:

\[
Y = H^H \cdot Q (H \cdot X + N_I) = H^H \cdot F (H \cdot X) + H^H \cdot N_O
\]

(133)

Or in scalar form:

\[
y_k = \sum_{m=1}^{M} h_k (m)^* \cdot \left( F \left( \sum_{k=1}^{K} h_k (m) \cdot x_k \right) + n_O (m) \right)
\]

(134)

where \( N_O \) is the vector of equivalent additive noise defined by (9), and \( F () \) is the elementwise equivalent transfer function calculated according to (6), (7) and (55).

The equivalent additive noise \( n_O (m) \) does not necessarily have a Gaussian distribution; however, if the number of ADCs, \( M \), is sufficiently large, then the resulting post-MRC additive noise can be approximated as Gaussian. Therefore, an equivalent ML decoder can be approximated as,

\[
\hat{X} = \arg \min_{\hat{X}} \left( (Y - \hat{Y} (\hat{X}))^H \cdot R \cdot (Y - \hat{Y} (\hat{X})) \right)
\]

(135)

where \( \hat{Y} (\hat{X}) \) is a lookup table representing NLD-aware MRC results precalculated for all \( N_{QAM}^K \) input possibilities,

\[
\hat{Y} (\hat{X}) = H^T \cdot F (H \cdot \hat{X})
\]

(136)

Or in scalar form:

\[
y_k = \sum_{m=0}^{M-1} h_k (m)^* \cdot \left( \sum_{k=1}^{K} h_k (m) \cdot x_k \right)
\]

(137)

Each element of this table is a vector of NLD-aware MRC output predictions, namely, \( F (H \cdot \hat{X}) \) for a particular \( \hat{X} \). Since this algorithm operates on MRC outputs, its complexity is almost identical to that of the naïve ML receiver (129), the only difference being the calculation of the NLD-aware MRC lookup table.

3) SIMULATION RESULTS

To confirm our equivalent model, we provide the following test. We simulate QAM64 uplink communication between a single user and a MIMO base station. The number of receiver antennas and the ADC resolution are simulation parameters. We assume a planar MIMO array with a line-of-sight (LoS) channel that is given by (117), where the angle of arrival \( \alpha_0 \) is a random variable with a uniform distribution from \(-\pi\) to \(+\pi\).

Apply UE input SNR to the \( k \)-th user as the output SNR of an ideal MRC receiver:

\[
\text{SNR}_k = \sum_{m=0}^{M-1} | h_k (m) |^2 \cdot \sigma_X^2 / \sigma_N^2
\]

(138)

We simulate the following receivers:

- For an array equipped with ideal floating point ADCs (no quantization error), we use the conventional ML receiver (128) that is optimal for infinite ADC resolution.
- For an array equipped with low-resolution ADCs:
  - The same conventional ML receiver as in the case of ideal ADCs above but is now a naïve one since it ignores the quantization effect (128).
  - Brute force ML receiver that incorporates NLD (130).
  - CEM-based ML receiver (135).

Figure 21 presents 64 MRC output realizations \( Y \) obtained from simulation and constellation points predicted by the lookup table (136) \( \hat{Y} (\hat{X}) \) when the angle of arrival \( \alpha = \pi / 12 \), the cumulative (sum over all antennas) input SNR is equal to 30 dB, MIMO array size is 1024 antennas and ADC resolution is 1 bit. From this figure, we can observe the effect of the ADC NLD and that MRC prediction matches the actual MRC realizations quite well.
A. Molev-Shteiman et al.: New Equivalent Model of a Quantizer With Noisy Input and Its Applications

Nonlinear MRC output constellation realizations for QAM16 are observed that are similar to those in [30]; however, we believe that the CEM model provides an efficient analytical way to estimate the points.

Figures 22 and 23 present, respectively, the resulting BER as a function of the cumulative input SNR for an array of 1024 antennas, each equipped with a pair of 1-bit ADCs, and for an array of 32 antennas, each equipped with a pair of 3-bit ADCs.

From these figures, we can make the following conclusions:
- For a low input SNR, NLD is small enough and the conventional ML receiver that ignores ADC nonlinearity works quite well.
- For a moderate to high SNR, the NLD-aware equivalent ML receiver significantly extends the SNR range of reasonable performance.

There is slight performance degradation of the NLD-aware equivalent ML receiver (135) relative to the brute force ML receiver (130) due to the Gaussian approximation of the post-MRC additive noise.

B. ITERATIVE MMSE RECEIVER

1) EXISTING MMSE RECEIVER
An ‘ADC-aware’ MMSE receiver estimates users’ signal vector $X$ as:

$$\hat{X} = R_{SS}^{-1} \cdot R_{SX} \cdot Q(S_I + N_I)$$  

(139)

where:

$$R_{SS} = E \left[ Q(S_I + N_I)H \cdot Q(S_I + N_I) \right]$$  

(140)

$$R_{SX} = E \left[ Q(S_I + N_I)H \cdot X \right]$$  

(141)

A ‘naïve’ MMSE receiver that does not consider the effect of ADCs uses the approximation:

$$R_{SS} \approx E \left[ (S_I + N_I)H \cdot (S_I + N_I) \right] = H^H \cdot H + I \cdot \sigma_N^2$$  

(142)

$$R_{SX} \approx E \left[ (S_I + N_I)H \cdot X \right] = H^H$$  

(143)

The analytical methods of the calculation of matrices $R_{SS}$ and $R_{SX}$ from channel estimation $H$ that considers the effect of ADC were presented in [32]. However, they have high complexity. Theoretically, we can also directly measure these matrices with a training sequence. However, the length of this sequence makes this approach impractical.

2) CEM-BASED ITERATIVE MMSE RECEIVER
According to CEM (17), we can write the uplink MIMO system equation (49) as:

$$S_O = Q(H \cdot X + N_I) = F(H \cdot X) + N_O$$  

(144)
Or in scalar form:
\[
s_O (m) = F \left( \sum_{k=1}^{K} h_k (m) \cdot \hat{x}_k \right) + n_O (m) \quad (145)
\]

We propose an iterative approach here. Let \( \hat{X}_0 \) denote the initial estimation of the vector \( X \) with elements \( \hat{x}_{k,0}, k = 0, \ldots, K - 1 \). Let us assume that from the \( i \)-th iteration, we obtain a sufficiently accurate estimation of each user signal \( \hat{x}_{k,i} \) with a sufficiently small error \( \Delta x_{k,i} \):
\[
x_k = \hat{x}_{k,i} + \Delta x_{k,i} \quad \text{and} \quad |\Delta x_{k,i}| \ll |\hat{x}_{k,i}| \quad (146)
\]

Then, the estimation of the desired signal \( \hat{s}_{l,i} (m) = \sum_{k=1}^{K} h_k (m) \cdot \hat{x}_{k,i} \) should also be accurate enough to allow approximation:
\[
F (s_l (m)) \approx F (\hat{s}_{l,i} (m)) + d (\hat{s}_{l,i} (m)) \cdot \Delta s_{l,i} (m) \quad (147)
\]

where \( \Delta s_{l,i} (m) \) represents the desired signal estimation error, and \( d (s) \) denotes the derivative of \( F (s) \) with respect to \( s \).
\[
\Delta s_{l,i} (m) = s_l (m) - \hat{s}_{l,i} (m) = \sum_{k=1}^{K} h_k (m) \cdot \Delta x_{k,i} \quad (148)
\]
\[
d (x) = \partial F (x) / \partial x = \sum_{r=1}^{R} q_r \cdot \partial \Pr (q_r | x) / \partial x \quad (149)
\]

For Gaussian input noise, the derivative is equal to:
\[
d (x) = q_1 \cdot \left( \exp \left( -\left( q_1 + (\Delta / 2) - x \right)^2 / \sigma_N^2 \right) / \sqrt{\pi \cdot \sigma_N^2} \right)
- q_R \cdot \left( \exp \left( -\left( q_R - (\Delta / 2) - x \right)^2 / \sigma_N^2 \right) / \sqrt{\pi \cdot \sigma_N^2} \right)
+ \sum_{r=2}^{R-1} q_r \cdot \left( \exp \left( -\left( q_r + (\Delta / 2) - x \right)^2 / \sigma_N^2 \right) / \sqrt{\pi \cdot \sigma_N^2} \right)
- \sum_{r=2}^{R-1} q_r \cdot \left( \exp \left( -\left( q_r - (\Delta / 2) - x \right)^2 / \sigma_N^2 \right) / \sqrt{\pi \cdot \sigma_N^2} \right) \quad (150)
\]

Then, we can approximate the uplink MIMO system equation (49) after the \( i \)-th iteration as:
\[
s_O (m) \approx F \left( \sum_{k=1}^{K} h_k (m) \cdot \hat{x}_{k,i} \right)
+ d \left( \sum_{k=1}^{K} h_k (m) \cdot \hat{x}_{k,i} \right) \cdot \sum_{k=1}^{K} h_k (m) \cdot \Delta x_{k,i} + n_O (m) \quad (151)
\]

Or in matrix form:
\[
S_O = F \left( H \cdot \hat{X}_i \right) + D \left( H \cdot \hat{X}_i \right) \cdot H \cdot \Delta X_i + N_O \quad (152)
\]

where \( \hat{X}_i \) and \( \Delta X_i \) denote vectors with length \( K \) and elements \( \hat{x}_k \) and \( \Delta x_{k,i} \), respectively, after the \( i \)-th iteration, \( N_O \) is a vector of length \( M \) with elements \( n_O (m) \), and \( D \left( H \cdot \hat{X}_i \right) \) is an \( M \) by \( M \) diagonal matrix with the \( m \)-th diagonal elements \( d (m, m) \) equal to:
\[
d (m, m) = d \left( \sum_{k=1}^{K} h_k (m) \cdot \hat{x}_{k,i} \right) \quad (153)
\]

Then, for next iteration, we can apply the MMSE approach to estimate vector \( \Delta X_i \) and to correct the initial estimation of vector \( \hat{X}_i \):
\[
\hat{X}_{i+1} = \hat{X}_i + \alpha \cdot \Delta X_i \quad (154)
\]

where \( \alpha \) is the iteration step size.
\[
\Delta X_i = \text{MMSE} \left( H \cdot \hat{X}_i \right) \cdot \left( S_O - F \left( H \cdot \hat{X}_i \right) \right) \quad (155)
\]

Given that \( N_O \) is a vector of uncorrelated components and is uncorrelated with \( X \), the MMSE weight matrix as a function of the desired signal estimation vector \( \hat{s}_{l,i} = H \cdot \hat{X}_i \) is equal to:
\[
\text{MMSE} \left( \hat{s}_{l,i} \right) = \left( H_i^H \cdot H_i + \frac{\sigma_N^2}{\sigma_X^2} \cdot I \right)^{-1} \cdot H_i^H \quad (156)
\]

where:
\[
H_i = D \left( \hat{s}_{l,i} \right) \cdot H \quad (157)
\]

At the first iteration, we assume \( \hat{X}_0 \) is zero which makes the MMSE matrix equal to that of the naive MIMO receiver that does not consider the effect of low-resolution ADCs. We iterate the process until a user signal vector \( X \) estimation of sufficient quality is obtained.

3) SIMULATION RESULTS

To confirm the theoretical analysis, we simulate a MIMO uplink base station receiver that employs an array of 10 antennas with a pair of 3-bit ADCs on each antenna and another receiver that employs an array of 1000 antennas with a pair of 1-bit ADCs on each antenna. We simulate both single-user and 4-user scenarios. The channel from each user to the base station is LOS, as described by (117). The transmit powers of all users are equal. For the single-user case, the angle of arrival is 0. For the 4-user case, the angles of arrival are \( \pi / 8, 3 \cdot \pi / 8, 5 \cdot \pi / 8 \) and \( 7 \cdot \pi / 8 \). The modulation is OFDM with QAM16 on each subcarrier. We employ the naive MMSE receiver that does not consider low-resolution ADC and the MMSE receiver with 1 iteration of NLD suppression (154) (\( \alpha = 1 \)). For reference, we also provide curves of the ideal MMSE receiver equipped with ideal ADC (no quantization) and the ideal MMSE for quantized input with measured ADC-aware matrices \( R_{SS} \) and \( \mathbf{R}_{SX} \). We directly measure the two matrices with a sufficiently long training sequence in order to minimize estimation errors (1000 OFDM symbols, or 10,240,000 samples). Each simulation point is a result of averaging bit errors over 10 independent OFDM symbols and all users. The simulation results are presented in figures 24, 25, 26 and 27.
The results again demonstrate that when the cumulative input SNR (post-MRC input SNR) is low and nonlinear distortion is mild, increasing the input SNR improves the performance and even the naïve MMSE receiver achieves good performance. However, after the input SNR exceeds a certain threshold, nonlinear distortion becomes dominant, and further increases in the cumulative input SNR cause significant performance degradation for a naïve receiver.

The 3-dB cumulative SNR thresholds shown in figure 18 again match well with the ‘knees’ on the BER curve for the naïve MMSE receivers. According to figure 18, the 3-dB cumulative SNR threshold is equal to 21 dB for both simulation cases.

Applying CEM-based NLD suppression techniques significantly extends the MIMO receiver operating range while maintaining linear complexity. The performance of the CEM-based MMSE receiver with iterative NLD suppression is only slightly worse than that of the brute force MMSE receiver taking the quantization effect into full consideration. The lower complexity is largely because the CEM equivalent transfer function takes only the desired signal as input, which is precisely the part of the input we aim to equalize, without the complicating effort of the input noise.

VII. CONCLUSION

We present a novel equivalent model for ADC and DAC quantizers that is uniquely suited for applications where the quantizer input consists of the desired signal and noise (noisy inputs). We believe the applications of the new model are wide-ranging, as most communication and sensing signals are the desired signal corrupted by noise. Here, we provide several examples of such applications.

- Design of digital dither that mitigates the effect of quantization error in the downlink MIMO transmitter and significantly reduces requirements regarding its DAC resolution.
- Determination of the minimal ADC resolution required for operation of a conventional (naïve) uplink MIMO receiver that does not consider the effect of a low-resolution ADC without significant performance degradation.
- Low-complexity ML and iterative MMSE receivers that consider the effect of a low-resolution ADC. It was shown that such algorithms can significantly increase the...
SNR operating range of a MIMO receiver with insufficient ADC resolution.

**APPENDIX A**

**EQUIVALENT OUTPUT NOISE PROPERTIES**

**Property (10):**

\[ E[n_o(m)] = 0 \]

*Proof:* From the output noise definition (9), it follows that the real and imaginary parts of the equivalent output noise \( n_o(m) \) are functions of the real and imaginary parts of the desired signal \( s_l(m) \) and the input noise \( n_l(m) \), respectively. The real and imaginary parts of both the desired signal \( s_l(m) \) and input noise \( n_l(m) \) are independent; therefore, the real and imaginary parts of the equivalent output noise \( n_o(m) \) are also independent.

From property (10), it follows that:

\[ E[\text{Re}(n_o(m))] = E[\text{Im}(n_o(m))] = 0 \]  

Therefore:

\[ E[\text{Re}(n_o(m)) \cdot \text{Im}(n_o(m))] = E[\text{Re}(n_o(m))] \cdot E[\text{Im}(n_o(m))] = 0 \]  

**Q.E.D.**

**Property (11):**

\[ E[\text{Re}(n_o(m)) \cdot \text{Im}(n_o(m))] = 0 \]

*Proof:* From the output noise definition (9), it follows that the real and imaginary parts of the equivalent output noise \( n_o(m) \) are functions of the real and imaginary parts of the desired signal \( s_l(m) \) and the input noise \( n_l(m) \), respectively. The real and imaginary parts of both the desired signal \( s_l(m) \) and input noise \( n_l(m) \) are independent; therefore, the real and imaginary parts of the equivalent output noise \( n_o(m) \) are also independent.

If a conditional expectation always equals zero, then the unconditional expectation is also zero.

**APPENDIX B**

**WORST-CASE NLD THEOREM**

For any complex random process \( w(m) \) with zero mean and variance \( \sigma_w^2 \):

\[ \sigma_R^2 = E\left[ \sum_{m=1}^{M} h(m)^* \cdot w(m) \right]^2 \leq \left( \sum_{m=1}^{M} |h(m)| \right)^2 \cdot \sigma_w^2 \]  

*Proof:*

\[ \sigma_R^2 = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} |h(m_1)| \cdot |h(m_2)| \cdot R_W(m_1, m_2) \]  

\[ R_W(m_1, m_2) = E[w(m_1)^* \cdot w(m_2)] \]  

\[ |R_W(m_1, m_2)| \leq \sigma_w^2 \]  

The \( \sigma_R^2 \) is maximal when all elements of the sum are maximal. Therefore:

\[ \max_{R_W} \left( \frac{\sigma_R^2}{\sigma_w^2} \right) = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} |h(m_1)| \cdot |h(m_2)| \cdot \max (|R_W(m_1, m_2)|) \]

\[ = \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} |h(m_1)| \cdot |h(m_2)| \cdot \sigma_w^2 \]

\[ = \left( \sum_{m=1}^{M} |h(m)| \right)^2 \cdot \sigma_w^2 \]  

**Q.E.D.**
APPENDIX C

WORST-CASE CHANNEL THEOREM

\[
\max_H \left( \frac{\left( \sum_{m=1}^{M} |h(m)| \right)^2}{\sum_{m=1}^{M} |h(m)|^2} \right) = M \tag{171}
\]

Proof: Let us denote:

\[
D_k (H) \triangleq \left( \frac{\sum_{m=k}^{M} |h(m)|}{\sum_{m=1}^{M} |h(m)|} \right)^2
\]

By taking the derivative \( \partial D_k / \partial |h(k)| \), we will find the maximum of \( D \) over \( |h(1)| \):

\[
\max_{|h(k)|} (D_k (H)) = 1 + D_{k+1} (H) \tag{173}
\]

Then:

\[
\max \left( \frac{\sum_{m=1}^{M} |h(m)|}{\sum_{m=1}^{M} |h(m)|^2} \right) = \max (D_1 (H)) = 1 + \max (D_2 (H)) = 2 + \max (D_2 (H)) = \ldots = M \tag{174}
\]

Q.E.D.

APPENDIX D

WORST-CASE SCENARIO EXAMPLE

An example realizing the worst-case NLD and channel described in Appendix B and C is the single-user MRC or MRT signal:

\[
s_I (m) = h(m)^* \cdot x \tag{175}
\]

where the channel is LoS:

\[
h(m) = g \cdot \exp \left( j \cdot \pi \cdot \sin (\alpha) \cdot m \right) = g \cdot (\pm 1 \text{ or } \pm j) \tag{176}
\]

It occurs when the angle of arrival/departure satisfies:

\[
\alpha = \left[ 0 \text{ or } \pm \pi / 2 \text{ or } \pm \pi / 3 \right] \tag{177}
\]

Proof: The NLD that generates signal \( s_I (m) \) is equal to:

\[
w(m) = F (s_I (m)) = g_O \cdot s_I (m) \tag{178}
\]

where \( g_O \) is the Bussgang-Rowe decomposition gain, and \( F (s) \) can be the equivalent ADC transfer function in the case of a MIMO receiver or a conventional DAC (without dither) transfer function in the case of a MIMO transmitter.

In both cases, according to (1), (3), (4) and (6), \( F (s) \) has properties:

\[
F (-s) = -F (s) \quad \text{and} \quad F (j \cdot s) = j \cdot F (s) \tag{179}
\]

Therefore, according to (176):

\[
w(m) = F (s_I (m)) - g_O \cdot s_I (m) = (h(m)/h(1))^* \times (F (h(1)^* \cdot x) - g_O \cdot h(1)^* \cdot x)
\]

\[
= h(m)^* \cdot w(1)/h(1)^* \tag{180}
\]

The post-MRC/MRT variance of such NLD reaches the maximal value defined by (176):

\[
\sigma_R^2 = E \left[ \frac{\sum_{m=1}^{M} h(m)^* \cdot w(m)^2}{\sum_{m=0}^{M-1} |h(m)|^2} \right]
\]

\[
= \sum_{m=1}^{M} |h(m)|^2 \cdot E \left[ |w(1)/h(1)^*|^2 \right]
\]

\[
= \sum_{m=1}^{M} |h(m)|^2 \cdot \sigma_w^2 \tag{181}
\]

For channel (176), function (171) reaches the maximum:

\[
M \left( \sum_{m=0}^{M-1} |h(m)|^2 \right) \left( \sum_{m=0}^{M-1} |h(m)|^2 \right) = M \tag{182}
\]

Q.E.D.

REFERENCES

[1] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, “Massive MIMO for next generation wireless systems,” IEEE Commun. Mag., vol. 52, no. 2, pp. 186–195, Feb. 2014.
[2] E. Björnson, M. Matthaiou, and M. Debbah, “Massive MIMO with non-ideal arbitrary arrays: Hardware scaling laws and circuit-aware design,” IEEE Trans. Wireless Comm., vol. 14, no. 8, pp. 4353–4368, Aug. 2015.
[3] R. H. Walden, “Analog-to-digital converter survey and analysis,” IEEE J. Sel. Areas Commun., vol. 17, no. 4, pp. 539–550, Apr. 1999.
[4] A. Gersho, Vector Quantization and Signal Compression. Cham, Switzerland: Springer, 2011.
[5] J. J. Bussgang, “Cross-correlation functions of amplitude-distorted Gaussian signals,” RLE, vol. 216, 1-14, Mar. 1952.
[6] H. E. Rowe, “Memoryless nonlinearities with Gaussian inputs: Elementary results,” Bell Syst. Tech. J., vol. 61, no. 7, pp. 1519–1526, Sep. 1982.
[7] J. Patel and K. Read, Handbook of the Normal Distribution, 2nd ed. Boca Raton, FL, USA: CRC Press, 1996.
[8] A. Puglielli, A. Townley, G. LaCaille, V. Milovanovi, P. Lu, K. Trotskovsky, A. Whitcombe, N. Narevsky, G. Wright, T. Courtade, E. Alon, B. Nikoli, and A. M. Niknejad, “Design of energy and cost-efficient massive MIMO arrays,” Proc. IEEE, vol. 104, no. 3, pp. 586–606, Mar. 2016.
[9] J. Zhang, L. Dai, X. Li, Y. Liu, and L. Hanzo, “On low-resolution ADCs in practical 5G millimeter-wave massive MIMO systems,” IEEE Commun. Mag., vol. 56, no. 7, pp. 205–211, Jul. 2018.
[10] O. Orhan, E. Erikp, and S. Rangan, “Low power analog-to-digital conversion in millimeter wave systems: Impact of resolution and bandwidth on performance,” in Proc. Inf. Theory Appl. Workshop (ITA), Feb. 2015, pp. 191–198.
[11] J. Choi, B. L. Evans, and A. Gatherer, “Resolution-adaptive hybrid MIMO architectures for millimeter wave communications,” IEEE Trans. Signal Process., vol. 65, no. 23, pp. 6201–6216, Dec. 2017.
[12] Q. Bai and J. A. Nossek, “Improving energy-efficiency of multi-antenna receivers via adaptation of ADC resolutions,” in Proc. 20th Eur. Wireless Conf., 2014, pp. 1–6.
[13] A. Kaushik, C. Tsinos, E. Vlachos, and J. Thompson, “Energy efficient ADC bit allocation and hybrid combining for millimeter wave MIMO systems,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Dec. 2019, pp. 1–6.
[14] J. Zhang, “On the spectral efficiency of massive MIMO systems with low-resolution ADCs,” IEEE Comm. Lett., vol. 20, no. 5, pp. 842–845, May 2016.
[15] Y. Dong and L. Qiu, “Spectral efficiency of massive MIMO systems with low-resolution ADCs and MMSE receiver,” IEEE Comm. Lett., vol. 21, no. 8, pp. 1771–1774, Aug. 2017.
ARKADY MOLEV-SHTEIMAN (Member, IEEE) received the B.S. and M.S. degrees in electrical and computer engineering from the Ben-Gurion University of the Negev, Israel, in 1995 and 2001, respectively. In 1995, he worked in leading positions for numerous Israeli startups that were acquired by Infineon, Intel, and Broadcom. He was relocated to the U.S. via startup acquisition, in 2006. He joined Futurewei Technologies, Inc., as a Principal Researcher, in 2015. His research interests include digital signal processing for communications and location estimation, especially in the context of massive MIMO. He is the author of 46 granted patents in this field.

XIAO-FENG QI (Senior Member, IEEE) received the Ph.D. degree from the University of Connecticut, in 1993. He has been the Senior Director of Radio Algorithms Research, Futurewei Technologies, Inc., since 2014, where he oversees a research portfolio of next-generation wireless communication and sensing technologies. He has been with AT&T Microelectronics, Globespan Semiconductors Inc., and Level One Communications, where he participated in delivering various wireless modem chips. In 2004, he took on design and supervisory roles in wireless product research and development, among them the Senior Director of advanced research and development at Intel, and the Technical Manager of the Cellular Systems Group, Broadcom. His current research interests include statistical signal processing techniques applied to wireless communications and sensing. He is the (co)inventor of more than 20 granted U.S. patents.

LAURENCE MAILANDER (Member, IEEE) received the Ph.D. degree from UC Santa Barbara, in 1996. He is currently a Senior Staff Scientist with Blue Danube Systems, Inc. He has previously worked in commercial and government research and development at Bell Labs, LGS Innovations LLC, Silvus Technologies, Inc., and Futurewei Technologies, Inc. He developed a practical equalizer that was adopted into Lucent’s UMTS product line for Rel 8. His research interests include MIMO systems, space-time equalization, and geolocation. He has served as the Track Chair of VTC and a Guest Editor for EURASIP.

BERTRAN M. HOCHWALD (Fellow, IEEE) was born in New York, NY, USA. He received the bachelor’s degree from the Swarthmore College, Swarthmore, PA, USA, the M.S. degree in electrical engineering from Duke University, Durham, NC, USA, and the M.A. degree in statistics and the Ph.D. degree in electrical engineering from Yale University, New Haven, CT, USA. From 1986 to 1989, he was with the Department of Defense, Fort Meade, MD, USA. He was a Research Associate and a Visiting Assistant Professor with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Champaign, IL, USA. In 1996, he joined the Mathematics of Communications Research Department, Bell Laboratories, Lucent Technologies, Inc., Murray Hill, NJ, USA, where he was a Distinguished Member of the Technical Staff. In 2005, he joined Beceem Communications, Santa Clara, CA, USA, as the Chief Scientist and the Vice-President of systems engineering. He was a Consulting Professor of electrical engineering with Stanford University, Palo Alto, CA, USA. In 2011, he joined the University of Notre Dame, Notre Dame, IN, USA, as a Freimann Professor of electrical engineering. He has 46 patents and has co-invented several well-known multiple-antenna techniques, including a differential method, linear dispersion codes, and multiuser vector perturbation methods. He is a Fellow of the National Academy of Inventors. He received several achievement awards while employed at the Department of Defense and the Prize Teaching Fellowship at Yale University. He received the 2006 Stephen O. Rice Prize for the Best Paper published in the IEEE TRANSACTIONS ON COMMUNICATIONS. He has coauthored a article that won the 2016 Best Paper Award by a young author in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS. He also won the 2018 H. A. Wheeler Prize Paper Award from the IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION. His Ph.D. students have won various honors for their Ph.D. research, including the 2018 Paul Baran Young Scholar Award from the Marconi Society. He has been listed as a Thomson Reuters Most Influential Scientific Mind in multiple years. He has served as an Editor for several IEEE journals and has given plenary and invited talks on various aspects of signal processing and communications.

NARAYAN PRASAD received the B.Tech. degree in electrical engineering from IIT Delhi, New Delhi, India, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Colorado Boulder, USA, in 2004. He was a Senior Researcher with NEC Laboratories America Inc., Princeton, NJ, USA. He is currently a Principal Researcher with Futurewei Technologies, Inc., Bridgewater, NJ, USA. His research interests include multiuser space-time communications, information theory, and radio resource management. He serves as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.