Channel Estimation in Broadband Millimeter Wave MIMO Systems With Few-Bit ADCs

Jianhua Mo, *Member, IEEE*, Philip Schniter ⃝, *Fellow, IEEE*, and Robert W. Heath, Jr. ⃝, *Fellow, IEEE*

Abstract—We develop a broadband channel estimation algorithm for millimeter wave (mmWave) multiple input multiple output (MIMO) systems with few-bit analog-to-digital converters (ADCs). Our methodology exploits the joint sparsity of the mmWave MIMO channel in the angle and delay domains. We formulate the estimation problem as a noisy quantized compressed-sensing problem and solve it using efficient approximate message passing (AMP) algorithms. In particular, we model the angle-delay coefficients using a Bernoulli–Gaussian-mixture distribution with unknown parameters and use the expectation-maximization forms of the generalized AMP and vector AMP algorithms to simultaneously learn the distributional parameters and compute approximately minimum mean-squared error (MSE) estimates of the channel coefficients. We design a training sequence that allows fast, fast Fourier transform based implementation of these algorithms while minimizing peak-to-average power ratio at the transmitter, making our methods scale efficiently to large numbers of antenna elements and delays. We present the results of a detailed simulation study that compares our algorithms to several benchmarks. Our study investigates the effect of SNR, training length, training type, ADC resolution, and runtime on channel estimation MSE, mutual information, and achievable rate. It shows that, in a mmWave MIMO system, the methods we propose to exploit joint angle-delay sparsity allow 1-bit ADCs to perform comparably to infinite-bit ADCs at low SNR, and 4-bit ADCs to perform comparably to infinite-bit ADCs at medium SNR.

Index Terms—Low resolution analog-to-digital converter, millimeter wave, channel estimation, approximate message passing.

I. INTRODUCTION

MILLIMETER wave (mmWave) communication is a promising technology for future outdoor cellular systems due to its potential to use very high bandwidth channels [2]. But larger bandwidths place difficult demands on the re-

ceiver's analog-to-digital converters (ADCs). For example, at rates above 100 Msamples per second, ADC power consumption increases quadratically with sampling frequency [3]. High-precision ADCs (e.g., ≥ 6 bits) with bandwidths sufficient for mmWave systems (e.g., ≥ 1 Gsamples/s) are either unavailable or may be too costly and power-hungry for portable devices [4]. One possible solution is to employ low-resolution ADCs, which enjoy low power consumption and simple hardware implementations. Low resolution ADCs have further benefits in MIMO systems, where a large number of ADCs are needed when digital baseband processing of all antenna outputs is desired. For example, in massive MIMO systems, it has been suggested to equip the base station with dozens of antennas and the same number of 1-bit ADCs [5]–[7].

In this paper, we consider a receiver architecture based on few-bit (i.e., 1-4 bit) ADCs, which act to quantize the inphase and quadrature baseband received signals. The achievable rate of the quantized MIMO channel was studied in [8]–[10] assuming channel-state information at the receiver (CSIR) was perfect but channel-state information at the transmitter (CSIT) was absent. In the latter case, and assuming equal transmission power at each antenna, [9] showed that QPSK is the optimum signaling strategy at low SNR. The perfect-CSIT case was studied in our previous work [11], where constellation design methods were proposed to maximize the achievable rate. These methods achieve much higher rates than QPSK signaling, especially at high SNR, but require CSIT.

Due to the nonlinear nature of quantization, channel estimation with few-bit ADCs is challenging. To estimate broadband SISO channel coefficients, the work [12], [13] proposed to transmit periodic bursty training sequences, dither the ADCs, and estimate each tap separately. MIMO channel estimation is even more challenging because the linear *combination* of transmitted signals from different antennas is quantized. In [14] and [15], the MIMO channel was estimated using least squares (LS) methods. In particular, the quantization error was treated as additive white Gaussian noise. As a result, a large estimation error was introduced.

To take into account quantization effects, iterative channel estimation methods using Expectation Maximization (EM) were proposed and analyzed in [1], [16], [17]. The proposed methods have high complexity, however, since each EM iteration computes a matrix inverse and many iterations are needed for convergence. In addition, [16], [17] considered the MIMO channel with small antenna arrays in the lower frequency UHF (ultra high frequency) band, and thus did not take into account the sparsity of mmWave channels [1], [18]–[20].

In recent work [6], [21]–[23], Approximate Message Passing (AMP) algorithms were used for channel estimation and (or)

symbol detection in the massive MIMO channel with few-bit ADCs. These works assumed that the channel coefficients follow an IID Gaussian distribution and did not exploit the sparsity inherent in mmWave channels. (Delay-domain sparsity was exploited in [23] but not angle-domain sparsity.) Also, since they were based on the massive-MIMO channel, these works did not exploit structures present in the broadband mmWave model that can be exploited for significant complexity reduction.

In this paper, we propose high-performance AMP-based channel estimation schemes for broadband mmWave MIMO channels with few-bit ADCs. The main contributions of our work are summarized as follows.

- We formulate the problem of estimating broadband mmWave channels under few-bit ADCs as a noisy, quantized, compressed-sensing problem. By leveraging sparsity in both the angle and delay domains, the massive MIMO channel can be accurately estimated by efficient algorithms and with relatively short training sequences. To our knowledge, our work is the first to take this approach.
- For channel estimation, we consider two AMP algorithms: the Generalized AMP (GAMP) algorithm from [24] and the Vector AMP (VAMP) algorithm from [25], [26]. For both, we use EM extensions [27], [28] to avoid the need to specify a detailed prior on the channel distribution. The GAMP and VAMP algorithms provide nearly minimum-MSE estimates with relatively low complexity in large-scale estimation problems and are therefore suitable for estimating broadband massive MIMO channels.
- We propose a novel training sequence design that results in low channel-estimation error, low complexity, and low peak-to-average power ratio. Low complexity is achieved through a novel FFT-based implementation that will be described in the sequel.
- We undertake a detailed experimental study of important design choices, such as ADC precision, the type and length of training sequence, and the type of estimation algorithm. When evaluating algorithms, we consider both performance and complexity. For performance, we consider several metrics: mean-square error (MSE), mutual information, and achievable rate.

From our experimental study, our main findings are as follows.

- Compared to other algorithms of which we are aware, the EM-VAMP algorithm has a superior performance-complexity tradeoff.
- Our FFT-based implementation facilitates low-complexity estimation of channels with large antenna numbers (e.g., 64×64) and delay spreads (e.g., 16 symbol intervals).
- Relative to infinite-bit ADCs, 1-bit ADCs incur only small performance losses at low SNR, and 3-4-bit ADCs incur only small losses up to medium SNRs.
- The MSEs of the EM-GAMP and EM-VAMP algorithms decay exponentially with training length, and the achievable rate is usually maximized by sending a relatively short training sequence.

A simple version of our proposed methodology was first published in [1]. Relative to [1], our current work expands from one-bit to few-bit ADCs, expands from narrowband to broadband channels, considers leakage effects in sparse-channel modeling, considers four training sequence designs instead of one, considers the VAMP algorithm as well as the GAMP al-
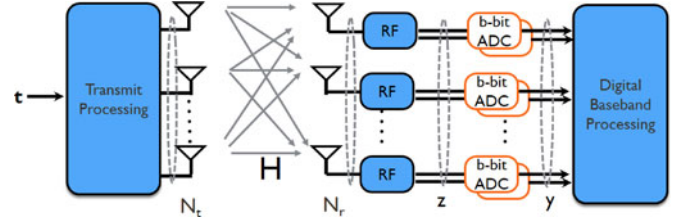

Fig. 1. A $N_r \times N_t$ MIMO system with one-bit quantization at the receiver. For each receiver antenna, there are two few-bit ADCs. Note that there is no limitation on the structure of the transmitter.

gorithm, proposes an FFT-based implementation that facilitates many more antennas and delays, considers Gaussian-mixture (GM) as well as Bernoulli-Gaussian priors, and incorporates EM learning of the prior parameters. Also, relative to [1], the experimental study in our current work is much more elaborate.

The paper is organized as follows. In Section II, we describe the broadband MIMO system model with few-bit ADCs. The channel characteristics of mmWave communications are then summarized in Section III. In Section IV, we present our channel-estimation algorithms and training-sequence designs. Simulation results are presented in Section V, followed by conclusions in Section VI.

Notation : a is a scalar, $\mathbf{a}$ is a vector and $\mathbf{A}$ is a matrix. $\mathrm{tr}(\mathbf{A})$, $\mathbf{A}^{\mathsf{T}}$, $\mathbf{A}^*$ and $||\mathbf{A}||_F$ represent the trace, transpose, conjugate transpose and Frobenius norm of a matrix $\mathbf{A}$. $\mathbf{A} \otimes \mathbf{B}$ denotes the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$. $\mathrm{vec}(\mathbf{A})$ is a vector stacking all the columns of $\mathbf{A}$. $\mathrm{unvec}(\cdot)$ is the opposite operation of $\mathrm{vec}(\cdot)$. $\mathcal{CN}(x; \mu, \sigma^2) \triangleq \frac{1}{\pi \sigma^2} \exp(-|x - \mu|^2/\sigma^2)$ is the probability density function of the circularly symmetric complex-Gaussian distribution with mean μ and variance σ^2. Finally, $\mathrm{j} \triangleq \sqrt{-1}$.

II. BROADBAND MIMO WITH FEW-BIT ADC

We consider a MIMO system with few-bit ADCs, as illustrated in Fig. 1. The transmitter is equipped with N_t antennas and the receiver is equipped with N_r antennas. A total of $2N_r$ few-bit ADCs separately quantize the real and imaginary parts of the received signal of each antenna. Assuming that the delay spread of the channel is limited to L symbol intervals and that carrier and symbol synchronization have already been performed, the quantizer output $\mathbf{y}[i] \in \mathbb{C}^{N_r \times 1}$ at time i can be written as

$$\mathbf{y}[i] = \mathcal{Q}\left(\sum_{\ell=0}^{L-1} \mathbf{H}[\ell]\mathbf{t}[i-\ell] + \mathbf{w}[i]\right), \quad (1)$$

where $\mathbf{H}[\ell] \in \mathbb{C}^{N_r \times N_t}$ is the baseband channel impulse response at lag ℓ, $\mathbf{t}[i] \in \mathbb{C}^{N_t \times 1}$ is the transmitted symbol at time i with average transmit power $\mathbb{E}\left[\mathbf{t}[i]^*\mathbf{t}[i]\right] = P_t$, and $\mathbf{w}[i] \sim \mathcal{CN}(\mathbf{0}, \sigma_w^2 \mathbf{I}_{N_r})$ is additive white Gaussian noise. Furthermore, $\mathcal{Q}(\cdot)$ denotes the quantization function, which is applied component-wise and separately to the real and imaginary parts.

In this paper, we assume that uniform mid-rise quantization is used. In particular, if x is a complex-valued scalar x, then

Resolution b	1-bit	2-bit	3-bit	4-bit	5-bit	6-bit	7-bit	8-bit
Stepsize Δ_b	$\sqrt{\frac{8}{\pi}}$ (≈ 1.5958)	0.9957	0.586	0.3352	0.1881	0.1041	0.0569	0.0308
NMSE η_b	$\frac{2-\pi}{\pi}$ (≈ 0.3634)	0.1188	0.03744	0.01154	0.003504	0.001035	0.0002999	0.00008543
SQNR (dB)	$10\log_{10}\frac{\pi}{2-\pi}$ (≈ 4.4)	9.25	14.27	19.38	24.55	29.85	35.23	40.68

$y = \mathcal{Q}(x)$ means

$$
y = \text{sign}\left(\text{Re}(x)\right)\left(\min\left(\left\lceil\frac{|\text{Re}(x)|}{\Delta_{\text{Re}}}\right\rceil, 2^{b-1}\right) - \frac{1}{2}\right)\Delta_{\text{Re}}
$$

$$
+ j\,\text{sign}\left(\text{Im}(x)\right)\left(\min\left(\left\lceil\frac{|\text{Im}(x)|}{\Delta_{\text{Im}}}\right\rceil, 2^{b-1}\right) - \frac{1}{2}\right)\Delta_{\text{Im}},
\tag{2}
$$

where $\Delta_{\text{Re}} \triangleq \left(\mathbb{E}\left[|\text{Re}(x)|^2\right]\right)^{\frac{1}{2}}\Delta_b$ and $\Delta_{\text{Im}} \triangleq \left(\mathbb{E}[|\text{Im}(x)|^2]\right)^{\frac{1}{2}}\Delta_b$, and where Δ_b is a stepsize that will be discussed in the sequel. In the special case of one-bit quantization, (2) becomes

$$
y = \text{sign}\left(\text{Re}(x)\right)\sqrt{\frac{2}{\pi}}\left(\mathbb{E}\left[|\text{Re}(x)|^2\right]\right)^{\frac{1}{2}}
$$

$$
+ j\,\text{sign}\left(\text{Im}(x)\right)\sqrt{\frac{2}{\pi}}\left(\mathbb{E}\left[|\text{Im}(x)|^2\right]\right)^{\frac{1}{2}}.
\tag{3}
$$

The average powers $\mathbb{E}\left[|\text{Re}(x)|^2\right]$ and $\mathbb{E}\left[|\text{Im}(x)|^2\right]$ can be easily measured by analog circuits before the ADC, as in automatic gain control (AGC). In this paper, x is circularly symmetric, and so $\mathbb{E}\left[|\text{Re}(x)|^2\right] = \mathbb{E}\left[|\text{Im}(x)|^2\right] = \frac{1}{2}\mathbb{E}\left[|x|^2\right]$, implying that $\Delta_{\text{Re}} = \Delta_{\text{Im}}$.

The quantization stepsize Δ_b is usually chosen to minimize the quantization MSE assuming a Gaussian input signal (see, e.g., [29]). These values of Δ_b are given in Table I assuming a unit-power input. The normalized MSE (NMSE), defined as $\eta_b \triangleq \mathbb{E}\left[|\mathcal{Q}(x) - x|^2\right]/\mathbb{E}\left[|x|^2\right]$, and the signal to quantization noise ratio (SQNR), defined as $10\log_{10}\frac{1}{\eta_b}$, are also listed in Table I. From the table it can be seen that $\Delta_b \sim 2^{-b}$, $\eta_b \sim 2^{-2b}$ and SQNR $\approx 5b$ dB. Notice that 4-bit ADCs yield quantization noise power around 20 dB below the signal power, which suggests that increasing ADC resolution beyond 4 bits should yield negligible performance improvement in the low and medium SNR regimes. This intuition will be verified by our simulations in Section V.

III. SPARSITY OF THE MMWAVE CHANNEL MODEL

In this section, we present the mmWave channel model assumed in our paper. The characteristics of the mmWave channel, especially the sparsity in the angle-delay domain, will be exploited by our proposed channel estimation algorithm.

A. Clustered MIMO Channel Model

The mmWave channel can be modeled using N_{cl} multipath clusters, where the nth cluster comprises N_{path}^n paths [19]. For the mth path of the nth cluster, we use $\alpha_{n,m}$, $\tau_{n,m}$, $\varphi_{n,m}^{\text{r}}$ (or $\theta_{n,m}^{\text{r}}$), $\varphi_{n,m}^{\text{t}}$ (or $\theta_{n,m}^{\text{t}}$) to denote the complex gain, delay, azimuth (or zenith) angle of arrival, and azimuth (or zenith) angle of departure, respectively. Using these quantities, the channel impulse response from (1) can be written as

$$
\mathbf{H}[\ell] = \sum_{n=1}^{N_{\text{cl}}}\sum_{m=1}^{N_{\text{path}}^n}\alpha_{n,m}\mathbf{a}_{\text{r}}(\varphi_{n,m}^{\text{r}}, \theta_{n,m}^{\text{r}})\mathbf{a}_{\text{t}}^*(\varphi_{n,m}^{\text{t}}, \theta_{n,m}^{\text{t}})
$$

$$
\times\, p(\ell T - \tau_{n,m}), \quad 0 \le \ell < L,
\tag{4}
$$

where $p(t)$ includes the effects of pulse shaping and analog/digital filtering, and where $\mathbf{a}_{\text{r}}$ and $\mathbf{a}_{\text{t}}$ are the array response vectors of the receive and transmit antenna arrays, respectively.

We assume that the receive and transmit arrays are each configured as a uniform planar array (UPA). This assumption is quite common. For example, in the massive MIMO systems described in [30], [31], UPAs with more than 100 antennas were tested. In the mmWave 5G cellular system prototype described in [32], there is a 256-element (16×16) UPA array panel at the base station, and two sets of 1×16 UPA arrays in the top and bottom portions of the mobile phone.

B. Angle-Delay Representation

The MIMO channel coefficients $\mathbf{H}[\ell] \in \mathbb{C}^{N_{\text{r}} \times N_{\text{t}}}$ are expressed in what is known as the "antenna aperture domain." We find it convenient to instead work with "angle domain" coefficients $\mathbf{X}[\ell] \in \mathbb{C}^{N_{\text{r}} \times N_{\text{t}}}$, as proposed in [33]. The two representations are connected through

$$
\mathbf{H}[\ell] = \mathbf{B}_{N_{\text{r}}}\mathbf{X}[\ell]\mathbf{B}_{N_{\text{t}}}^*, \quad 0 \le \ell < L,
\tag{5}
$$

where $\mathbf{B}_{N_{\text{r}}} \in \mathbb{C}^{N_{\text{r}} \times N_{\text{r}}}$ and $\mathbf{B}_{N_{\text{t}}} \in \mathbb{C}^{N_{\text{t}} \times N_{\text{t}}}$ are the steering matrices for the transmitter and receiver arrays, respectively. Denote N_{r}^{e} (N_{r}^{a}) as the number of receive antennas in the elevation (azimuth) direction, and N_{t}^{e} (N_{t}^{a}) as the number of transmit antennas in the elevation (azimuth) direction. With an $N_{\text{r}}^{\text{e}} \times N_{\text{r}}^{\text{a}}$ receive UPA and an $N_{\text{t}}^{\text{e}} \times N_{\text{t}}^{\text{a}}$ transmit UPA, we have that $N_{\text{r}} = N_{\text{r}}^{\text{a}}N_{\text{r}}^{\text{e}}$, $N_{\text{t}} = N_{\text{t}}^{\text{a}}N_{\text{t}}^{\text{e}}$, and [34]

$$
\mathbf{H}[\ell] = \left(\mathbf{F}_{N_{\text{r}}^{\text{a}}} \otimes \mathbf{F}_{N_{\text{r}}^{\text{e}}}\right)\mathbf{X}[\ell]\left(\mathbf{F}_{N_{\text{t}}^{\text{a}}} \otimes \mathbf{F}_{N_{\text{t}}^{\text{e}}}\right)^*,
\tag{6}
$$

where $\mathbf{F}_{N_{\text{r}}^{\text{a}}} \in \mathbb{C}^{N_{\text{r}}^{\text{a}} \times N_{\text{r}}^{\text{a}}}$, $\mathbf{F}_{N_{\text{r}}^{\text{e}}} \in \mathbb{C}^{N_{\text{r}}^{\text{e}} \times N_{\text{r}}^{\text{e}}}$, $\mathbf{F}_{N_{\text{t}}^{\text{a}}} \in \mathbb{C}^{N_{\text{t}}^{\text{a}} \times N_{\text{t}}^{\text{a}}}$, and $\mathbf{F}_{N_{\text{t}}^{\text{e}}} \in \mathbb{C}^{N_{\text{t}}^{\text{e}} \times N_{\text{t}}^{\text{e}}}$ are unitary Discrete Fourier Transform (DFT) matrices. The (i, j)th entry of the matrix $\mathbf{X}[\ell]$ can be interpreted as the channel gain between the jth discrete transmit angle and the ith discrete receive angle [33].

According to the measurement results reported in [19], the number of clusters N_{cl} tends to be relatively few in the mmWave band as compared to lower-frequency bands. Also, the number of antenna elements used tends to be large in order to counteract the effects of path loss, which is much more severe in the mmWave band as compared to low-frequency bands. Hence, in mmWave applications, the number of channel clusters is usually much fewer than the number of scalar coefficients in $\{\mathbf{H}[\ell]\}_{l=0}^{L-1}$ or $\{\mathbf{X}[\ell]\}_{l=0}^{L-1}$, i.e., $N_{\text{cl}} \ll N_{\text{r}}N_{\text{t}}L$.
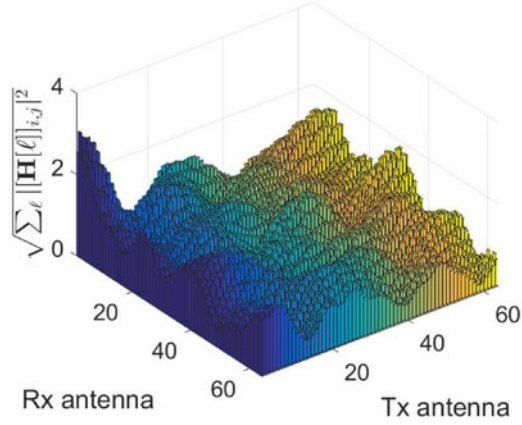
(a) Antenna domain channel
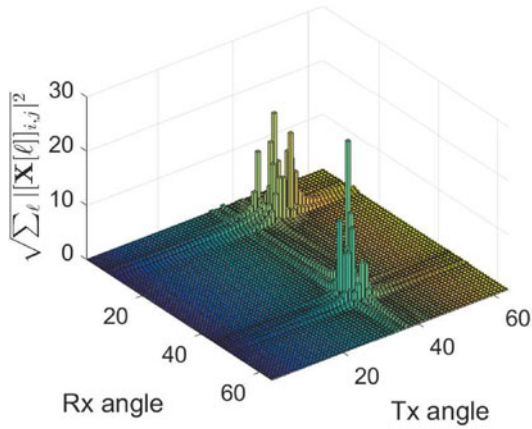
(b) Angle domain channel

Fig. 2.   An example of a broadband MIMO channel with $N_{\rm cl} = 2$ multipath clusters, as seen through a link with 8×8 UPAs at both ends. Subfigure (a) plots in linear scale the mean-squared coefficient magnitude in the antenna aperture domain $\sqrt{\sum_\ell |[\mathbf{H}[\ell]]_{i,j}|^2}$, and subfigure (b) plots in linear scale the same for the angle domain $\sqrt{\sum_\ell |[\mathbf{X}[\ell]]_{i,j}|^2}$. The angle domain representation in (b) clearly shows two groups of large-magnitude entries, each of which corresponds to one multipath cluster.

To show precisely how channel sparsity manifests in the angle domain, Fig. 2 provides an example of a simulated broadband mmWave channel with delay spread $L = 16$ and $N_{\rm cl} = 2$ multipath clusters, each consisting of 10 paths and 7.5 degrees of azimuth and elevation angular spread, as seen through 8×8 UPAs at both ends of the link (i.e., $N_{\rm t} = N_{\rm r} = 64$) and raised-cosine pulse-shape filtering with roll-off factor $= 0$. These parameters follow the urban macro (UMa) NLOS channel measurements at 28 GHz reported in [19]. Fig. 2(a) plots the mean-squared coefficient magnitude in the antenna aperture domain, i.e., $\sqrt{\sum_\ell |[\mathbf{H}[\ell]]_{i,j}|^2}$, while subfigure (b) plots the mean-squared coefficient magnitude in the angle domain, i.e., $\sqrt{\sum_\ell |[\mathbf{X}[\ell]]_{i,j}|^2}$. Fig. 2(b) shows that, in the angle domain, the channel energy is concentrated in two locations, each corresponding to one multipath cluster.

We emphasize that the angle-delay channel $\{\mathbf{X}[\ell]\}_{\ell=0}^{L-1}$ is *jointly* sparse in the angle and delay domains. This joint sparsity is illustrated in Fig. 3(a) for a 4×16 MIMO channel with delay spread $L = 16$ (see the details in Section V). However,

$\{\mathbf{X}[\ell]\}_{\ell=0}^{L-1}$ is *not exactly sparse*, exhibiting what is known as leakage, in that none of its coefficients are expected to be exactly zero valued. The approximate nature of the sparsity is also visible in Fig. 3(a). Our approach based on GAMP will be robust to leakage effects.

IV. PROPOSED CHANNEL ESTIMATION ALGORITHM

Motivated by the channel model given in the last section, we now develop an efficient algorithm to estimate the approximately sparse angle-delay domain channel from few-bit measurements and a known training sequence.

A. Problem Formulation

We assume that the training consists of a block transmission of length $N_{\rm p}$, denoted as $\mathbf{T} \in \mathbb{C}^{N_{\rm t} \times N_{\rm p}}$, with a cyclic prefix of length L. After discarding the cyclic prefix, the measurements take the form

$$\mathbf{Y} = \mathcal{Q}(\mathbf{Z} + \mathbf{W}), \tag{7}$$

where $\mathbf{W}$ is additive Gaussian noise and $\mathbf{Z} \in \mathbb{C}^{N_{\rm r} \times N_{\rm p}}$ is the unquantized noiseless received signal block, of the form

$$\mathbf{Z} = \sum_{\ell=0}^{L-1} \mathbf{H}[\ell] \mathbf{T} \mathbf{J}_\ell \tag{8}$$

$$= \sum_{\ell=0}^{L-1} \mathbf{B}_{N_{\rm r}} \mathbf{X}[\ell] \mathbf{B}_{N_{\rm t}}^* \mathbf{T} \mathbf{J}_\ell \tag{9}$$

$$= \mathbf{B}_{N_{\rm r}} \underbrace{\left[\mathbf{X}[0]\ \mathbf{X}[1]\ \mathbf{X}[2]\ \cdots\ \mathbf{X}[L-1] \right]}_{\triangleq\ \mathbf{X}}$$

$$\times \underbrace{\begin{bmatrix} \mathbf{B}_{N_{\rm t}}^* & & & \\ & \mathbf{B}_{N_{\rm t}}^* & & \\ & & \ddots & \\ & & & \mathbf{B}_{N_{\rm t}}^* \end{bmatrix}}_{=\ \mathbf{I}_L \otimes \mathbf{B}_{N_{\rm t}}^*} \underbrace{\begin{bmatrix} \mathbf{T}\mathbf{J}_0 \\ \mathbf{T}\mathbf{J}_1 \\ \vdots \\ \mathbf{T}\mathbf{J}_{L-1} \end{bmatrix}}_{\triangleq\ \widetilde{\mathbf{T}}} \tag{10}$$

$$= \mathbf{B}_{N_{\rm r}} \mathbf{X} \underbrace{\begin{bmatrix} \mathbf{B}_{N_{\rm t}}^* \mathbf{T}\mathbf{J}_0 \\ \mathbf{B}_{N_{\rm t}}^* \mathbf{T}\mathbf{J}_1 \\ \vdots \\ \mathbf{B}_{N_{\rm t}}^* \mathbf{T}\mathbf{J}_{L-1} \end{bmatrix}}_{\triangleq\ \mathbf{C}}, \tag{11}$$

where $\mathbf{J}_\ell \in \mathbb{R}^{N_{\rm p} \times N_{\rm p}}$ is the ℓ-circulant-delay matrix.

We now rewrite $\mathbf{Z}$ in vector form $\mathbf{z} \triangleq \mathrm{vec}(\mathbf{Z})$ as

$$\mathbf{z} = \left(\mathbf{C}^{\mathsf{T}} \otimes \mathbf{B}_{N_{\rm r}} \right) \mathrm{vec}(\mathbf{X}). \tag{12}$$

Defining $\mathbf{y} \triangleq \mathrm{vec}(\mathbf{Y})$, $\mathbf{x} \triangleq \mathrm{vec}(\mathbf{X})$, and $\mathbf{w} \triangleq \mathrm{vec}(\mathbf{W})$, the quantized noisy output becomes

$$\mathbf{y} = \mathcal{Q}\left(\underbrace{\left(\mathbf{C}^{\mathsf{T}} \otimes \mathbf{B}_{N_{\rm r}} \right)}_{\triangleq\ \mathbf{A}} \mathbf{x} + \mathbf{w} \right). \tag{13}$$

The channel-estimation problem reduces to the following: estimate the angle-delay channel coefficients $\mathbf{x} \in \mathbb{C}^{N_{\rm t} N_{\rm r} L}$ from the
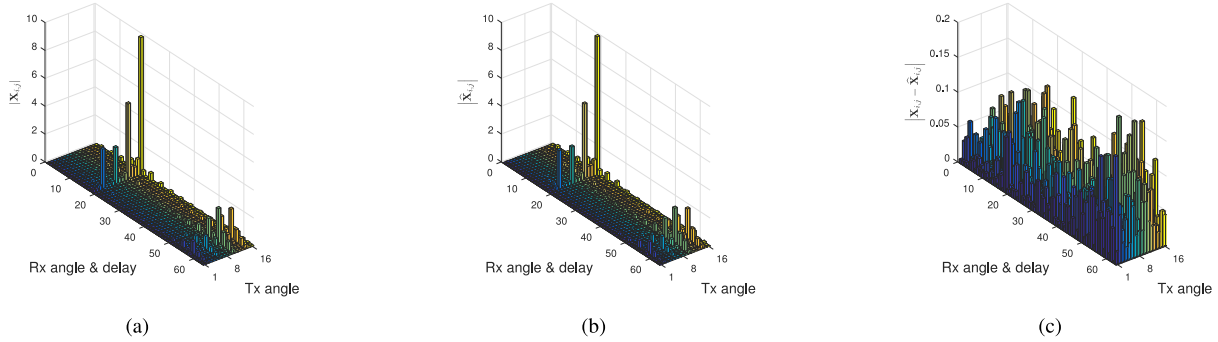
Fig. 3.   The true angle-delay channel magnitudes are plotted in (a), the magnitudes of the EM-GM-VAMP estimate in (b), and the estimation error magnitudes in (c) in linear scale. In particular, (a) shows $|[\mathbf{X}[\ell]]_{i,j}|$ at the (x, y) location $x = 4\ell + i$ and $y = j$. For this example, there were 2 clusters in a 4×16 MIMO channel with a delay spread of 16 samples. The block length of the training signal was 512 and the SNR was 10 dB. The ADC had 4-bit resolution.

noisy quantized received signal $\mathbf{y} \in \mathbb{C}^{N_{\mathrm{r}} N_{\mathrm{p}} \times 1}$ under the known linear transform $\mathbf{A} \in \mathbb{C}^{N_{\mathrm{r}} N_{\mathrm{p}} \times N_{\mathrm{t}} N_{\mathrm{r}} L}$. For use in the sequel, we define $N_y \triangleq N_{\mathrm{r}} N_{\mathrm{p}}$ and $N_x \triangleq N_{\mathrm{t}} N_{\mathrm{r}} L$.

Because the angle-delay channel is sparse (approximately due to leakage), this channel estimation problem can be viewed as an instance of *noisy quantized compressed sensing*. Several methods have been proposed for noisy quantized compressed sensing, including quantized iterative hard thresholding (QIHT) [35], [36], convex relaxation [37], [38], and GAMP [39], [40]. It has also been proposed to treat the quantization error as if it were additive white Gaussian noise [14], [15]. All of these methods are strongly dependent on the sparsity rate assumed for the channel, either explicitly or through the specification of a regularization term or prior distribution. In practice, though, the the channel sparsity rate, $1 - \lambda_0$ in (17) and (18), is a priori unknown.

B. EM-AMP Algorithms

For channel estimation, we propose to modify two approaches that combine expectation maximization (EM) with AMP as a means of avoiding the need to specify a prior. The first is based on a combination of generalized AMP (GAMP) [24] and EM, which was proposed in [27] but, to our knowledge, has never been applied to noisy few-bit quantized compressive sensing. The second is based on a combination of the recently proposed vector AMP (VAMP) [25], [26] and EM [28]. To our knowledge, neither VAMP nor EM-VAMP have been applied to noisy few-bit quantized compressive sensing. We focus on these AMP approaches because they offer nearly minimum MSE (MMSE) performance while being computationally efficient. We provide background on these methods here to make our paper self-contained.

1) EM-GAMP: Suppose that a random vector $\mathbf{x}$ with IID components $x_i \sim p_X$ is linearly transformed to produce $\mathbf{z} = \mathbf{A}\mathbf{x}$, which then propagates through a probabilistic measurement channel $p(\mathbf{y}|\mathbf{z}) = \prod_i p_{Y|Z}(y_i|z_i)$. Given knowledge of $\mathbf{y}$, $\mathbf{A}$, p_X, and $p_{Y|Z}$, we would like to compute the MMSE estimate of $\mathbf{x}$. The GAMP algorithm [24] approaches this computationally difficult problem through a sequence of simple scalar estimation problems and matrix multiplies. Remarkably, when $\mathbf{A}$ is very large with IID (sub)Gaussian entries, the behavior of GAMP is rigorously characterized by a scalar state evolution [24], [41]. When this state evolution has a unique fixed point, GAMP converges to the MMSE solution. In practice, $\mathbf{A}$ may

---

Algorithm 1: The EM-GAMP Algorithm.

1: **define:**

$$p_{Z|Y,P}(z_i|y_i, \widehat{p}_i; \nu_p) \triangleq \frac{p_{Y|Z}(y_i|z_i)\mathcal{CN}(z_i; \widehat{p}_i, \nu_p)}{\int_z p_{Y|Z}(y_i|z)\mathcal{CN}(z; \widehat{p}_i, \nu_p)} \quad (14)$$

$$p_{X|R}(x_i|\widehat{r}_i; \nu_r, \boldsymbol{\theta}) \triangleq \frac{p_X(x_i; \boldsymbol{\theta})\mathcal{CN}(x_i; \widehat{r}_i, \nu_r)}{\int_x p_X(x; \boldsymbol{\theta})\mathcal{CN}(x; \widehat{r}_i, \nu_r)} \quad (15)$$

2: **initialize:** $\widehat{\mathbf{s}} = \mathbf{0}$, $\boldsymbol{\theta}$, $\widehat{x}_i = \int_x x\, p_X(x; \boldsymbol{\theta})$ $\forall i$,
$\quad \nu_x = \int_x |x - \widehat{x}_1|^2 p_X(x; \boldsymbol{\theta})$,
3: **for** $k = 0, 1, \ldots, N_{\max}$ **do**
4: $\quad \nu_p \leftarrow \|\mathbf{A}\|_F^2 \nu_x / N_y$
5: $\quad \widehat{\mathbf{p}} \leftarrow \mathbf{A}\widehat{\mathbf{x}} - \nu_p \widehat{\mathbf{s}}$
6: $\quad \nu_z \leftarrow N_y^{-1} \sum_{i=1}^{N_y} \mathrm{Var}_{Z|Y,P}\left[z_i \mid y_i, \widehat{p}_i; \nu_p\right]$
7: $\quad \widehat{z}_i \leftarrow \mathbb{E}_{Z|Y,P}\left[z_i \mid y_i, \widehat{p}_i; \nu_p\right]$, $\quad \forall i$
8: $\quad \nu_s \leftarrow \left(1 - \nu_z/\nu_p\right)/\nu_p$
9: $\quad \widehat{s}_i \leftarrow \nu_p^{-1}\left(\widehat{z}_i - \widehat{p}_i\right)$, $\quad \forall i$
10: $\quad \nu_r^{-1} \leftarrow \|\mathbf{A}\|_F^2 \nu_s / N_x$
11: $\quad \widehat{\mathbf{r}} \leftarrow \widehat{\mathbf{x}} + \nu_r \mathbf{A}^* \widehat{\mathbf{s}}$
12: $\quad \nu_x \leftarrow N_x^{-1} \sum_{i=1}^{N_x} \mathrm{Var}_{X|R}\left[x_i \mid \widehat{r}_i; \nu_r, \boldsymbol{\theta}\right]$
13: $\quad \widehat{x}_i \leftarrow \mathbb{E}_{X|R}\left[x_i \mid \widehat{r}_i; \nu_r, \boldsymbol{\theta}\right]$, $\quad \forall i$
14: $\quad$ update the parameters $\boldsymbol{\theta}$ using EM algorithm;
15: **end for**
16: **return** $\widehat{\mathbf{x}}$.

---

not be a very large IID (sub)Gaussian matrix, in which case the theoretical guarantees of GAMP do not hold. Still, the estimates it provides after a few (e.g., < 25) iterations are often very close to MMSE (see, e.g., [27]).

GAMP requires specification of p_X, which is unknown in practice. To circumvent this problem, [27] proposed to approximate the true p_X by a Gaussian-mixture with parameters $\boldsymbol{\theta}$ learned by an EM algorithm. In the E-step of the EM algorithm, GAMP's posterior approximation is used in place of the true posterior, which is NP-hard to compute. The resulting EM-GAMP algorithm was empirically analyzed in [27], where it was shown to give similar performance to true-p_X GAMP for a wide range of true p_X (e.g., sparse, heavy tailed, discrete). Generalizations of EM-GAMP were theoretically analyzed in [42] using the GAMP state evolution.

Algorithm 1 details the steps of EM-GAMP. Lines 1-13 are from the original GAMP algorithm and line 14 is the EM

Algorithm 2: The EM-VAMP Algorithm.

1: **define:** $p_{Z|Y,P}$ and $p_{X|R}$ from (14)-(15).
2: **initialize:** $\mathbf{r}_1, \mathbf{p}_1, \nu_1, \tau_1$ and $\boldsymbol{\theta}$.
3: **for** $k = 0, 1, \ldots, N_{\max}$ **do**
4:     // Scalar estimation of x_i
5:     $\widehat{x}_{1i} \leftarrow \mathbb{E}_{X|R}[x_i \,|\, r_{1i}; \nu_1, \boldsymbol{\theta}], \quad \forall i$
6:     $\alpha_1 \leftarrow \nu_1^{-1} N_x^{-1} \sum_{i=1}^{N_x} \mathrm{Var}_{X|R}[x_i \,|\, r_{1i}; \nu_1, \boldsymbol{\theta}]$
7:     $\mathbf{r}_2 \leftarrow (\widehat{\mathbf{x}}_1 - \alpha_1 \mathbf{r}_1)/(1 - \alpha_1), \quad \nu_2 \leftarrow \nu_1 \alpha_1/(1 - \alpha_1)$
8:     // Scalar estimation of z_i
9:     $\widehat{z}_{1i} \leftarrow \mathbb{E}_{Z|Y,P}[z_i \,|\, y_i, p_{1i}; \tau_1], \quad \forall i$
10:    $\beta_1 \leftarrow \tau_1^{-1} N_y^{-1} \sum_{i=1}^{N_y} \mathrm{Var}_{Z|Y,P}[z_i \,|\, y_i, p_{1i}; \tau_1]$
11:    $\mathbf{p}_2 \leftarrow (\widehat{\mathbf{z}}_1 - \beta_1 \mathbf{p}_1)/(1 - \beta_1), \quad \tau_2 \leftarrow \tau_1 \beta_1/(1 - \beta_1)$
12:    // LMMSE estimation of $\mathbf{x}$
13:    $\widehat{\mathbf{x}}_2 \leftarrow \mathbf{V}_A (\mathbf{S}_A^* \mathbf{S}_A \nu_2/\tau_2 + \mathbf{I})^{-1}(\mathbf{S}_A^* \mathbf{U}_A^* \mathbf{p}_2 \nu_2/\tau_2 + \mathbf{V}_A^* \mathbf{r}_2)$
14:    $\alpha_2 \leftarrow N_x^{-1} \sum_{n=1}^{N_x} \tau_2/(s_n^2 \nu_2 + \tau_2)$
15:    $\mathbf{r}_1 \leftarrow (\widehat{\mathbf{x}}_2 - \alpha_2 \mathbf{r}_2)/(1 - \alpha_2), \quad \nu_1 \leftarrow \nu_2 \alpha_2/(1 - \alpha_2)$
16:    // LMMSE estimation of $\mathbf{z}$
17:    $\widehat{\mathbf{z}}_2 \leftarrow \mathbf{A}\widehat{\mathbf{x}}_2$
18:    $\beta_2 \leftarrow (1 - \alpha_2)N_y/N_x$
19:    $\mathbf{p}_1 \leftarrow (\widehat{\mathbf{z}}_2 - \beta_2 \mathbf{p}_2)/(1 - \beta_2), \quad \tau_1 \leftarrow \tau_2 \beta_2/(1 - \beta_2)$
20:    update the parameters $\boldsymbol{\theta}$ using EM algorithm;
21: **end for**
22: **return** $\widehat{\mathbf{x}}_1$.

---

update. Lines 6-7 can be interpreted as computing the posterior mean and variance of z_i under the likelihood $p_{Y|Z}(y_i|z_i)$ and a pseudo-prior $z_i \sim \mathcal{CN}(\widehat{p}_i, \nu_p)$, where $\widehat{p}_i$ and ν_p are updated in lines 4-5. Likewise, lines 12-13 can be interpreted as computing the posterior mean and variance of x_i under the prior $p_X(x_i; \boldsymbol{\theta})$ and pseudo-measurement $\widehat{r}_i = x_i + \mathcal{CN}(0, \nu_r)$, where $\widehat{r}_i$ and ν_r are updated in lines 8-11. We note that, if $\mathbf{A}$ has a fast (e.g., FFT-based) implementation, then EM-GAMP can leverage this in lines 5 and 11. For more details on EM-GAMP, we refer the reader to [24], [27].

2) EM-VAMP: The VAMP algorithm [25], [26] aims to solve exactly the same problem targeted as GAMP and—like GAMP—is rigorously characterized by a scalar state-evolution. But VAMP's state evolution holds for a much broader class of matrices $\mathbf{A}$: those that are right-rotationally invariant (RRI). For $\mathbf{A}$ to be RRI, its SVD $\mathbf{A} = \mathbf{U}_A \mathbf{S}_A \mathbf{V}_A^*$ should have $\mathbf{V}_A$ drawn uniformly from the set of unitary matrices; there is no restriction on $\mathbf{U}_A$ and $\mathbf{S}_A$. For MMSE estimation, VAMP must be given the prior p_X and likelihood $p_{Y|Z}$. When they are not available, approximations can be learned through the EM methodology, as described in [28].

Algorithm 2 details the steps of EM-VAMP. As can be seen, it alternates between nonlinear scalar estimations (lines 4-11), linear vector estimations (lines 12-19), and an EM update (line 20). For scalar estimation, the conditional mean and variance of x_i under prior $p_X(x_i)$ and pseudo-measurements $r_{1i} = x_i + \mathcal{CN}(0, \nu_1)$ are computed in lines 5-6, and the conditional mean and variance of z_i under likelihood $p_{Y|Z}(y_i|z_i)$ and pseudo-prior $z_i \sim \mathcal{CN}(p_{1i}, \tau_1)$ are computed in lines 9-10. Lines 13 and 17 then compute the joint MMSE estimate of $\mathbf{x}$

and $\mathbf{z}$ under the pseudo-prior

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{z} \end{bmatrix} \sim \mathcal{CN}\left(\begin{bmatrix} \mathbf{r}_2 \\ \mathbf{p}_2 \end{bmatrix}, \begin{bmatrix} \nu_2 \mathbf{I} & \\ & \tau_2 \mathbf{I} \end{bmatrix} \right) \tag{16}$$

and the constraint $\mathbf{z} = \mathbf{A}\mathbf{x}$. We note that EM-VAMP can leverage fast implementations of $\mathbf{U}_A$ and $\mathbf{V}_A$ if they exist. For more details, we refer the reader to [25], [26], [28].

C. EM-AMP Algorithms for mmWave Channel Estimation with Few-bit ADCs

To apply EM-GAMP and EM-VAMP to mmWave channel-estimation with few-bit ADCs, we choose the (approximating) prior family as either a Bernoulli Gaussian-mixture (GM) or Bernoulli-Gaussian (BG), with unknown parameters $\boldsymbol{\theta}$.[1] That is, the coefficients x_i of $\mathbf{x}$ are assumed to be drawn from one of the following:

$$\text{GM: } p_X(x_i; \boldsymbol{\theta}) = \lambda_0 \delta(x_i) + \sum_i \lambda_i \, \mathcal{CN}(x_i; \mu_i, \phi_i) \, \forall i, \tag{17}$$

$$\text{BG: } p_X(x_i; \boldsymbol{\theta}) = \lambda_0 \delta(x_i) + (1 - \lambda_0)\mathcal{CN}(x_i; 0, \phi) \, \forall i, \tag{18}$$

where $\lambda_0 = \mathrm{Prob}\{x = 0\}$ and $\{\lambda_i\}, \{\mu_i\}, \{\phi_i\}$ are the weights, means, and variances of the Gaussian mixture, respectively (which are all included in $\boldsymbol{\theta}$), and $\delta(\cdot)$ is the Dirac delta distribution. Since GM has more degrees of freedom than BG, it can form a better fit to the true channel distribution and thus leads to better estimation performance than BG. As shown in the simulations, the complexity of GM is a bit higher than BG since more parameters have to be estimated. The expressions for $\mathbb{E}_{X|R}[x_i \,|\, \widehat{r}_1; \nu_r, \boldsymbol{\theta}]$ and $\mathrm{Var}_{X|R}[x_i \,|\, \widehat{r}_i; \nu_r, \boldsymbol{\theta}]$ needed in lines 12-13 of Algorithm 1 and lines 5-6 of Algorithm 2 can be found in [27], as can the EM-update expressions for $\boldsymbol{\theta}$.

For the few-bit quantizer (2), the likelihood $p_{Y|Z}$ is

$$p_{Y|Z}(y_i|z_i) \triangleq \mathrm{Prob}\{y_i = \mathcal{Q}(z_i + w_i) \,|\, z_i\} \tag{19}$$

$$= \int_{w \in \mathcal{Q}^{-1}(y_i) - z_i} \mathcal{CN}(w; 0, \sigma_w^2). \tag{20}$$

The expressions for $\mathbb{E}_{Z|Y,P}[z_i \,|\, y_i, \widehat{p}_i; \nu_p]$ and $\mathrm{Var}_{Z|Y,P}[z_i \,|\, y_i, \widehat{p}_i; \nu_p]$ needed in lines 6-7 of Algorithm 1 and lines 9-10 of Algorithm 2 can be obtained by following the procedures in [43, Chapter 3.9]. Further details can be found in [6, Appendix A].

D. Computational Issues and Training Sequence Design

In practical mmWave applications, the dimensions $N_y = N_r N_p$ and $N_x = N_t N_r L$ of the matrix $\mathbf{A}$ are expected to be very large. For example, in our simulations, we consider $N_t = N_r = 64$, $N_p = 1024$, and $L = 16$, which yield $N_y = N_x = 65536$. Storing such an $\mathbf{A}$ as an explicit matrix using 4-bytes each for the real and imaginary components would require 32 GB of memory, which is inconvenient in many applications. Even when $\mathbf{A}$ fits in memory, the computational complexity of EM-GAMP and EM-VAMP (or any known algorithm, for that matter) will be impractical if $\mathbf{A}$ is treated as an explicit matrix, due to, e.g., per-iteration matrix-vector multiplies with $\mathbf{A}$ and $\mathbf{A}^*$.

---

[1]In this paper, we model $\mathbf{x}$ using an IID sparse prior. We do not exploit possible correlation within $\mathbf{x}$, which may improve estimation accuracy.

For EM-GAMP, these problems are avoided if $\mathbf{A}$ and $\mathbf{A}^*$ can be represented as implicit fast operators. Likewise, for EM-VAMP, $\mathbf{U}_A$, $\mathbf{U}_A^*$, $\mathbf{V}_A^*$, and $\mathbf{V}_A$ should be fast operators. Recalling from (13) that $\mathbf{A} = \mathbf{C}^\mathsf{T} \otimes \mathbf{B}_{N_\mathrm{r}}$ with Fourier $\mathbf{B}_{N_\mathrm{r}}$ and $\mathbf{T}$-dependent $\mathbf{C}^\mathsf{T}$, we see that the training sequence $\mathbf{T}$ will determine whether $\mathbf{A}$ is a fast operator. Thus, the design of the training signal plays a vital role in the practical implementability of mmWave channel estimation.

There are, in fact, three considerations for the design of the training signal $\mathbf{T}$:

1) For GAMP to be computationally efficient, $\mathbf{A}$ and $\mathbf{A}^*$ should be fast operators, and for VAMP to be computationally efficient, $\mathbf{U}_A$, $\mathbf{U}_A^*$, $\mathbf{V}_A^*$, $\mathbf{V}_A$ should be.
2) For GAMP to converge to a good solution, $\mathbf{A}$ should be sufficiently dense and have sufficiently low peak-to-average squared-singular-value ratio [44].
3) To improve the efficiency of the power amplifier at the transmitter, the elements of $\mathbf{T}$ should have low peak-to-average power ratio (PAPR).

1) Training Structure: To satisfy these three considerations, we propose to structure $\mathbf{T} \in \mathbb{C}^{N_\mathrm{t} \times N_\mathrm{p}}$ as follows. Denote the first row of $\mathbf{T}$, i.e., the signal sent by the first transmit antenna, as $\mathbf{t}^\mathsf{T} \triangleq [t[0], \dots, t[N_\mathrm{p} - 1]]$. We fix N_p at an integer multiple of $N_\mathrm{t} L$ and construct $\mathbf{T}$ such that its nth row (for $n \geq 0$) is the nL-place circular shift of $\mathbf{t}^\mathsf{T}$. That is, each antenna sends a circularly shifted version of the signal transmitted by the first antenna. The elements of $\mathbf{T}$ (for $m \geq 0$) are then

$$[\mathbf{T}]_{n,m} = t[\langle m - nL \rangle_{N_\mathrm{p}}], \tag{21}$$

where $\langle \cdot \rangle_{N_\mathrm{p}}$ is the modulo-N_p remainder. With the construction in (21), $\mathbf{T}\mathbf{J}_0$ from (10) contains the $\{0, L, 2L, \dots, (N_\mathrm{t} - 1)L\}$ shifts of $\mathbf{t}$. Likewise, $\mathbf{T}\mathbf{J}_1$ contains the $\{1, L + 1, 2L + 1, \dots, (N_\mathrm{t} - 1)L + 1\}$ shifts of $\mathbf{t}$, and $\mathbf{T}\mathbf{J}_{L-1}$ contains the $\{L - 1, 2L - 1, 3L - 1, \dots, N_\mathrm{t}L - 1\}$ right shifts of $\mathbf{t}$. Altogether, the rows of $\widetilde{\mathbf{T}}$ in (10) will consist of the first $N_\mathrm{t}L$ circular shifts of the sequence $\mathbf{t}$. We can thus re-order the rows in $\widetilde{\mathbf{T}}$ to make it a Toeplitz matrix, which we will denote by $\overline{\mathbf{T}}$ in the sequel.

2) Fast Implementation of $\mathbf{A}\widehat{\mathbf{x}}$ and $\mathbf{A}^*\widehat{\mathbf{s}}$: Using the structure above, $\mathbf{A}\widehat{\mathbf{x}}$ and $\mathbf{A}^*\widehat{\mathbf{s}}$ can be efficiently computed. Notice

$$\text{unvec}\left(\mathbf{A}\widehat{\mathbf{x}}\right)$$

$$= \mathbf{B}_{N_\mathrm{r}} \widehat{\mathbf{X}} \left(\mathbf{I}_L \otimes \mathbf{B}_{N_\mathrm{t}}^*\right) \widetilde{\mathbf{T}} \tag{22}$$

$$\overset{(a)}{=} \mathbf{B}_{N_\mathrm{r}} \widehat{\mathbf{X}} \left(\mathbf{I}_L \otimes \mathbf{B}_{N_\mathrm{t}}^*\right) \mathbf{K}^{(L,N_\mathrm{t})} \overline{\mathbf{T}} \tag{23}$$

$$\overset{(b)}{=} \mathbf{B}_{N_\mathrm{r}} \widehat{\mathbf{X}} \mathbf{K}^{(L,N_\mathrm{t})} \mathbf{K}^{(N_\mathrm{t},L)} \left(\mathbf{I}_L \otimes \mathbf{B}_{N_\mathrm{t}}^*\right) \mathbf{K}^{(L,N_\mathrm{t})} \overline{\mathbf{T}} \tag{24}$$

$$\overset{(c)}{=} \mathbf{B}_{N_\mathrm{r}} \widehat{\mathbf{X}} \mathbf{K}^{(L,N_\mathrm{t})} \left(\mathbf{B}_{N_\mathrm{t}}^* \otimes \mathbf{I}_L\right) \overline{\mathbf{T}} \tag{25}$$

$$\overset{(d)}{=} \mathbf{B}_{N_\mathrm{r}} \left[\overline{\mathbf{T}}^\mathsf{T} \left(\mathbf{B}_{N_\mathrm{t}}^* \otimes \mathbf{I}_L\right) \overline{\mathbf{X}}\right]^\mathsf{T}, \tag{26}$$

where (a) follows with commutation matrix[2] $\mathbf{K}^{(L,N_\mathrm{t})} \in \mathbb{R}^{N_\mathrm{t}L \times N_\mathrm{t}L}$; (b) follows from $\mathbf{K}^{(L,N_\mathrm{t})}\mathbf{K}^{(N_\mathrm{t},L)} = \mathbf{I}_{N_\mathrm{t}L}$; (c) follows from $\mathbf{K}^{(N_\mathrm{t},L)} \left(\mathbf{I}_L \otimes \mathbf{B}_{N_\mathrm{t}}^*\right) \mathbf{K}^{(L,N_\mathrm{t})} = \mathbf{B}_{N_\mathrm{t}}^* \otimes \mathbf{I}_L$; and (d)

follows from $\overline{\mathbf{X}} \triangleq (\widehat{\mathbf{X}}\mathbf{K}^{(L,N_\mathrm{t})})^\mathsf{T}$, $(\mathbf{A} \otimes \mathbf{B})^\mathsf{T} = \mathbf{A}^\mathsf{T} \otimes \mathbf{B}^\mathsf{T}$, and the symmetry of $\mathbf{B}_{N_\mathrm{t}}^*$. Note that $\widehat{\mathbf{X}}\mathbf{K}^{(L,N_\mathrm{t})}$ is merely a reordering of the columns in $\widehat{\mathbf{X}}$.

We now show that (26) has a fast implementation. First, for any $\mathbf{v} \in \mathbb{C}^{LN_\mathrm{t} \times 1}$, notice that

$$\left(\mathbf{B}_{N_\mathrm{t}}^* \otimes \mathbf{I}_L\right)\mathbf{v} = \text{vec}\left(\mathbf{V}\mathbf{B}_{N_\mathrm{t}}^*\right) = \text{vec}\left(\left(\mathbf{B}_{N_\mathrm{t}}\mathbf{V}^*\right)^*\right), \tag{27}$$

where $\mathbf{V} \in \mathbb{C}^{L \times N_\mathrm{t}}$ is the column-wise matricization of $\mathbf{v}$. Second, for any $\mathbf{u} \in \mathbb{C}^{N_\mathrm{t} \times 1}$, notice that

$$\mathbf{B}_{N_\mathrm{t}}\mathbf{u} = \left(\mathbf{F}_{N_\mathrm{t}^\mathrm{a}} \otimes \mathbf{F}_{N_\mathrm{t}^\mathrm{e}}\right)\mathbf{u} = \text{vec}\left(\mathbf{F}_{N_\mathrm{t}^\mathrm{e}} \mathbf{U} \mathbf{F}_{N_\mathrm{t}^\mathrm{a}}\right), \tag{28}$$

where $\mathbf{U} \in \mathbb{C}^{N_\mathrm{t}^\mathrm{e} \times N_\mathrm{t}^\mathrm{a}}$ is the column-wise matricization of $\mathbf{u}$. Therefore, each of the N_r columns of the multiplication $\left(\mathbf{B}_{N_\mathrm{t}}^* \otimes \mathbf{I}_L\right) \overline{\mathbf{X}}$ in (26) can be accomplished by N_t^e-point and N_t^a-point FFTs.[3] The result of that multiplication is then left-multiplied by $\overline{\mathbf{T}}^\mathsf{T}$, which can be performed via fast convolution using an N_p-point FFT (since $\overline{\mathbf{T}}^\mathsf{T}$ contains the first $N_\mathrm{t}L$ columns of an $N_\mathrm{p} \times N_\mathrm{p}$ circulant matrix). Finally, the left-multiplication by $\mathbf{B}_{N_\mathrm{r}}$ can be performed, as in (28), using N_r^a-point and N_r^e-point FFTs. In summary, there are a total of $N_\mathrm{r}L$ FFTs of length N_p, one inverse-FFT of length N_p, $N_\mathrm{t}^\mathrm{a}N_\mathrm{r}$ FFTs of length N_t^e, $N_\mathrm{t}^\mathrm{e}N_\mathrm{r}$ FFTs of length N_t^a, $N_\mathrm{r}^\mathrm{a}N_\mathrm{p}$ FFTs of length N_r^e, and $N_\mathrm{r}^\mathrm{e}N_\mathrm{p}$ FFTs of length N_r^a. Multiplications of the form $\mathbf{A}^*\widehat{\mathbf{s}}$ can be computed similarly. Note also that the memory footprint of $\mathbf{A}$ reduces to that of $\mathbf{t} \in \mathbb{C}^{N_\mathrm{p}}$.

3) Choice of the Training Sequence $\mathbf{t}$: For GAMP to work well, we want the measurement matrix $\mathbf{A}$ to have sufficiently low peak-to-average squared-singular-value ratio [44]. Likewise, for VAMP to be fast, we want the singular-vector matrices $\mathbf{U}_A$ and $\mathbf{V}_A$ of $\mathbf{A}$ to be fast operators. We now show that both concerns can be addressed through the design of $\mathbf{t}$.

From (10), (11), and (13), the matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \left(\left(\mathbf{I}_L \otimes \mathbf{B}_{N_\mathrm{t}}^*\right) \widetilde{\mathbf{T}}\right)^\mathsf{T} \otimes \mathbf{B}_{N_\mathrm{r}}. \tag{29}$$

Since $\mathbf{I}_L \otimes \mathbf{B}_{N_\mathrm{t}}^*$ and $\mathbf{B}_{N_\mathrm{r}}$ are unitary matrices, the singular values of $\mathbf{A}$ have the form

$$\underbrace{\lambda_1, \dots, \lambda_1}_{N_\mathrm{r}}, \underbrace{\lambda_2, \dots, \lambda_2}_{N_\mathrm{r}}, \dots, \underbrace{\lambda_{N_\mathrm{t}L}, \dots, \lambda_{N_\mathrm{t}L}}_{N_\mathrm{r}}, \tag{30}$$

where $\lambda_1, \lambda_2, \dots, \lambda_{N_\mathrm{t}L}$ are the singular values of $\widetilde{\mathbf{T}}$. Therefore, we want $\widetilde{\mathbf{T}}$ to have sufficiently low peak-to-average squared-singular-value ratio. One way to ensure this property is to choose $\mathbf{t}$ as a Zadoff-Chu (ZC) sequence [45].

A length-N_p ZC sequence is defined as

$$t[k] = \begin{cases} \sqrt{\dfrac{P_\mathrm{t}}{N_\mathrm{t}}} \exp\left(\mathrm{j}\dfrac{\pi k(k+1)}{N_\mathrm{p}}\right) & \text{if } N_\mathrm{p} \text{ is odd}, \\[2ex] \sqrt{\dfrac{P_\mathrm{t}}{N_\mathrm{t}}} \exp\left(\mathrm{j}\dfrac{\pi k^2}{N_\mathrm{p}}\right) & \text{if } N_\mathrm{p} \text{ is even}. \end{cases} \tag{31}$$

---

[2]The commutation matrix $\mathbf{K}^{(m,n)}$ is the $mn \times mn$ matrix which, for any $m \times n$ matrix $\mathbf{M}$, transforms $\text{vec}(\mathbf{M})$ into $\text{vec}(\mathbf{M}^\mathsf{T})$, i.e., $\mathbf{K}^{(m,n)}\text{vec}(\mathbf{M}) = \text{vec}(\mathbf{M}^\mathsf{T})$.

[3]To see this, note that each column of the multiplication $(\mathbf{B}_{N_\mathrm{t}}^* \otimes \mathbf{I}_L)\overline{\mathbf{X}}$ can be represented as (27) with appropriate $\mathbf{v}$, and thus computed via $\mathbf{B}_{N_\mathrm{t}}\mathbf{V}^*$ due to (27). Then, $\mathbf{B}_{N_\mathrm{t}}\mathbf{V}^*$ can be computed columnwise through subproblems of the form (28) with appropriate $\mathbf{u}$, which in turn can be tackled through N_t^e-point FFTs of the columns of $\mathbf{U}$ and N_t^a-point FFTs of the rows of $\mathbf{U} = \text{unvec}(\mathbf{u})$, as shown by the right side of (28).

The periodic autocorrelation of a ZC sequence equals a scaled Kronecker delta. As a result, $\widetilde{\mathbf{T}}\widetilde{\mathbf{T}}^* = \frac{P_t N_p}{N_t}\mathbf{I}_{N_t L}$, and so $\lambda_1 = \lambda_2 = \cdots = \lambda_{N_t L} = \sqrt{P_t N_p / N_t}$. Consequently, $\mathbf{A}$ will have the minimum possible peak-to-average squared-singular-value ratio, which is good for GAMP. Furthermore, this singular-value structure implies that the singular-vector matrices can be chosen as $\mathbf{V}_A = \mathbf{I}$ and $\mathbf{U}_A = \sqrt{\frac{N_t}{P_t N_p}}\mathbf{A}$. Since $\mathbf{V}_A$ and $\mathbf{U}_A$ have fast implementations, they are good for VAMP.

An additional benefit of choosing the ZC sequence for $\mathbf{t}$ is its constant-modulus property, which ensures that $\mathbf{T}$ has a low peak-to-average power ratio PAPR. For this reason, ZC sequences are currently used as reference signals in LTE [46].

4) Other Choices of $\mathbf{T}$: In Section V, we investigate other training designs, for example, $\mathbf{T}$ composed of Golay complementary sequences, IID QPSK entries, and IID Gaussian entries. Although their recovery performance is comparable to our proposed design, they lead to GAMP and VAMP algorithms with much higher complexity due to the lack of fast methods to compute the matrix-vector multiplications with $\mathbf{A}, \mathbf{A}^*, \mathbf{U}_A, \mathbf{U}_A^*, \mathbf{V}_A, \mathbf{V}_A^*$.

E. Benchmark Algorithms

In this section, we describe several other signal reconstruction approaches that we will use as benchmarks. The first approach is the LS estimator adopted in [5], [14], [15]. In this case

$$\widehat{\mathbf{X}}_{\text{LS}} = \mathbf{B}_{N_r}^* \mathbf{Y} \mathbf{C}^{\dagger}, \qquad (32)$$

where $\mathbf{C}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{C}$. When $\mathbf{C}$ has full row-rank, we have

$$\widehat{\mathbf{X}}_{\text{LS}} = \mathbf{B}_{N_r}^* \mathbf{Y} \mathbf{C}^* \left(\mathbf{C}\mathbf{C}^*\right)^{-1}. \qquad (33)$$

Another approach is obtained by linearizing the quantizer and applying LMMSE [47], [48]. We explain the linearization procedure because it will be used later for achievable-rate analysis. Using Bussgang's theorem, the quantizer output $\mathbf{y}$ can be decomposed into a signal component plus a distortion $\mathbf{w}_q$ that is uncorrelated with the signal component [10], i.e.,

$$\mathbf{y} = \mathcal{Q}\left(\mathbf{A}\mathbf{x} + \mathbf{w}\right)$$
$$= (1 - \eta_b)\left(\mathbf{A}\mathbf{x} + \mathbf{w}\right) + \mathbf{w}_q, \qquad (34)$$

where η_b is the NMSE given in Table I. Under additional mild assumptions, [10, eq. (30)] showed that the "effective" noise $\widehat{\mathbf{w}} \triangleq (1 - \eta_b)\mathbf{w} + \mathbf{w}_q$ in the linearized model

$$\mathbf{y} = (1 - \eta_b)\mathbf{A}\mathbf{x} + \widehat{\mathbf{w}} \qquad (35)$$

has covariance $\boldsymbol{\Sigma}_{\widehat{w}}$ that is well approximated as

$$\boldsymbol{\Sigma}_{\widehat{w}} \approx (1 - \eta_b)\Big((1 - \eta_b)\mathbb{E}\left[\mathbf{w}\mathbf{w}^*\right]$$
$$+ \eta_b \text{diag}\big(\mathbb{E}\left[\mathbf{A}\mathbf{x}\mathbf{x}^*\mathbf{A}^*\right] + \mathbb{E}\left[\mathbf{w}\mathbf{w}^*\right]\big)\Big) \triangleq \widehat{\boldsymbol{\Sigma}}_{\widehat{w}}, \qquad (36)$$

where $\text{diag}(\mathbf{M})$ is formed by zeroing the off-diagonal elements of $\mathbf{M}$. If we furthermore assume $\mathbb{E}[\mathbf{x}\mathbf{x}^*] = \sigma_x^2 \mathbf{I}$ with σ_x^2 given

in (45) then, due to the independence of $\mathbf{x}$ and $\mathbf{A}$,

$$\widehat{\boldsymbol{\Sigma}}_{\widehat{w}} = (1 - \eta_b)\sigma_w^2 \mathbf{I}_{N_r N_p} + \eta_b (1 - \eta_b)\sigma_x^2 \text{diag}\big(\mathbb{E}[\mathbf{A}\mathbf{A}^*]\big) \qquad (37)$$

$$\overset{(a)}{=} (1 - \eta_b)\sigma_w^2 \mathbf{I}_{N_r N_p} + \eta_b (1 - \eta_b)\sigma_x^2 \text{diag}\big(\mathbb{E}[\widetilde{\mathbf{T}}^*\widetilde{\mathbf{T}}]^{\mathsf{T}} \otimes \mathbf{I}_{N_r}\big)$$

$$= \big((1 - \eta_b)\sigma_w^2 + \eta_b (1 - \eta_b)P_t L \sigma_x^2\big)\mathbf{I}_{N_r N_p}, \qquad (38)$$

where (a) uses $\mathbf{A}\mathbf{A}^* = (\mathbf{C}^*\mathbf{C})^{\mathsf{T}} \otimes \mathbf{I}_{N_r} = (\widetilde{\mathbf{T}}^*\widetilde{\mathbf{T}})^{\mathsf{T}} \otimes \mathbf{I}_{N_r}$, which follows from from (13), the unitary property of $\mathbf{B}_{N_r}$, and (10)–(11), and (38) uses $\text{diag}(\mathbb{E}[\widetilde{\mathbf{T}}^*\widetilde{\mathbf{T}}]) = P_t L \mathbf{I}_{N_p}$, which follows from the transmitter power constraint. Thus, the effective noise $\widehat{\mathbf{w}}$ can be approximated as spectrally white with variance

$$\sigma_{\widehat{w}}^2 = (1 - \eta_b)\big(\sigma_w^2 + \eta_b P_t L \sigma_x^2\big). \qquad (39)$$

The approximation (36) has been shown to be quite accurate for MIMO communication, especially at low SNR [10], [49]. Note, however, that $\widehat{\mathbf{w}}$ is non-Gaussian.

Assuming $\mathbb{E}[\mathbf{x}\mathbf{x}^*] = \sigma_x^2 \mathbf{I}$ (as above) and leveraging the result (38) that $\mathbb{E}[\widehat{\mathbf{w}}\widehat{\mathbf{w}}^*] \approx \sigma_{\widehat{w}}^2 \mathbf{I}$, one can straightforwardly derive the linear MMSE (LMMSE) estimator of $\mathbf{x}$ from $\mathbf{y}$ in (35). We will refer to it as the "approximate LMMSE (ALMMSE) estimator" due to the approximation (36). Expressed in terms of $\mathbf{X} = \text{unvec}(\mathbf{x})$, it takes the form

$$\widehat{\mathbf{X}}_{\text{ALMMSE}}$$
$$= \mathbf{B}_{N_r}^* \mathbf{Y}\mathbf{C}^* \left((1 - \eta_b)\mathbf{C}\mathbf{C}^* + \left(\frac{\sigma_w^2}{\sigma_x^2} + \eta_b P_t L\right)\mathbf{I}_{N_t L}\right)^{-1}, \qquad (40)$$

which is similar to (33) but with a regularized inverse.

A third benchmark algorithm follows by applying sparse reconstruction to the linearized model (35), which—unlike the methods above—leverages the fact that $\mathbf{x}$ is approximately sparse due to leakage. In particular, we used the "SPGL1" algorithm [50] to solve the basis pursuit denoising (BPDN) problem

$$\arg\min \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y} - (1 - \eta_b)\mathbf{A}\mathbf{x}\|^2 \leq \sigma_{\widehat{w}}^2 N_p N_r, \qquad (41)$$

with $\sigma_{\widehat{w}}^2$ defined in (39).[4] Similar to EM-GAMP and EM-VAMP, the computational complexity of SPGL1 is dominated by matrix-vector multiplications with $\mathbf{A}$ and $\mathbf{A}^*$, and so the fast implementation (26) is used.

A fourth benchmark algorithm is the quantized iterative hard thresholding (QIHT) algorithm proposed in [35], [36]. At iteration $k = 1, 2, \ldots$, the estimate is updated as

$$\widehat{\mathbf{s}}^{(k)} = \mathcal{Q}\left(\mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}^{(k)}\right), \qquad (42)$$

$$\widehat{\mathbf{x}}^{(k+1)} = h_K\left(\widehat{\mathbf{x}}^{(k)} + \tau \mathbf{A}^*\widehat{\mathbf{s}}^{(k)}\right), \qquad (43)$$

where $h_K(\mathbf{x})$ is the hard thresholding operator that zeros all but the K largest (in magnitude) components of $\mathbf{x}$. For convergence, it is suggested that $\tau < \|\mathbf{A}\|_2^{-2}$ for spectral norm $\|\mathbf{A}\|_2$. Our simulations used $\widehat{\mathbf{x}}^{(1)} = \mathbf{0}$, $\tau = 0.1\frac{N_t}{P_t N_p}$, and $K = N_x/100$. That is, QIHT kept only the largest 1% of the elements in each iteration.

---

[4] We also tried the ℓ_1-based method from [37], but we experienced numerical problems with their first-order method. Since our $\mathbf{A}$ was too large to fit in memory, we could not use the CVX implementation of [37] nor [38].

F. Norm Estimation of the Channel

With a one-bit ADC, all amplitude information is lost during quantization. Therefore, it is difficult to precisely recover the channel norm $\|\mathbf{x}\|$, especially at high SNR. In previous work [6], [7], the variance of the channel was assumed to be known to avoid this issue. However, it is possible in practice to estimate the channel norm from the average signal power received by the antenna circuit before quantization, for example, the AGC circuit.

Let us use $\widehat{\mathbf{z}} \triangleq (\mathbf{C}^{\mathsf{T}} \otimes \mathbf{B}_{N_r})\mathbf{x} + \mathbf{w}$ to denote the noisy unquantized signal. We assume that it is possible to measure the total received power across all antennas, which is $\mathbb{E}[\|\widehat{\mathbf{z}}\|^2]$. Under the assumption that $\mathbb{E}[\mathbf{x}\mathbf{x}^*] = \sigma_x^2 \mathbf{I}$, it is straightforward to show that

$$\mathbb{E}[\|\widehat{\mathbf{z}}\|^2] = \|\mathbf{C}^{\mathsf{T}} \otimes \mathbf{B}_{N_r}\|_F^2 \sigma_x^2 + N_p N_r \sigma_w^2. \tag{44}$$

In the UPA case, we have $\|\mathbf{C}^{\mathsf{T}} \otimes \mathbf{B}_{N_r}\|_F^2 = N_r \|\mathbf{C}\|_F^2$, and for ZC training we have $\|\mathbf{C}\|_F^2 = P_t N_p L$, which implies that

$$\sigma_x^2 = \frac{\mathbb{E}[\|\widehat{\mathbf{z}}\|^2] - N_p N_r \sigma_w^2}{P_t L N_p N_r}. \tag{45}$$

Under $\mathbb{E}[\mathbf{x}\mathbf{x}^*] = \sigma_x^2 \mathbf{I}$, the law of large numbers then says that $\|\mathbf{x}\| \approx \sigma_x \sqrt{N_r N_t L}$ for large $N_r N_t L$. Thus, for all channel estimators, we normalized the channel estimate $\widehat{\mathbf{x}}$ as follows,

$$\widetilde{\mathbf{x}} = \widehat{\mathbf{x}}\frac{\widehat{\|\mathbf{x}\|}}{\|\widehat{\mathbf{x}}\|} \text{ for } \widehat{\|\mathbf{x}\|} \triangleq \sigma_x\sqrt{N_r N_t L}, \tag{46}$$

since it reduced the channel estimation error in all cases.

V. SIMULATION RESULTS

We first provide an illustrative example to visualize the behavior of the proposed channel estimation procedure. In Fig. 3, we show the result of estimating a broadband $N_r \times N_t = 4 \times 16$ MIMO channel. The transmitter was equipped with a 4×4 UPA and the receiver with a 2×2 UPA. The channel had $N_{cl} = 2$ clusters and delay spread of at most $L = 16$ samples. 4-bit ADCs and raised-cosine pulse-shaping with roll-off factor zero was used. The training length was shifted-ZC of length $N_p = 512$. The true magnitudes of the angle-delay channel are plotted in Fig. 3(a) while the estimated magnitudes are plotted in Fig. 3(b). As seen in the figure, the EM-GM-VAMP algorithm can estimate the channel $\mathbf{X}$ quite accurately: the two clusters can be easily identified in Fig. 3(b) and the estimation error in Fig. 3(c) is small. The joint angle-delay sparsity of the mmWave channel is visible in Fig. 3(a), as is the *approximate* nature of the sparsity.

In the following subsections, we consider a more practical scenario, where the transmitter and receiver are both equipped with 8×8 UPAs. We consider a channel with 4 clusters, each of 10 paths, with azimuth and elevation angular spread of 7.5 degrees. The delay spread was at most 16 samples. These numbers were chosen according to the urban macro (UMa) NLOS channel measurement results at 28 GHz, as given in the white paper [19]. The results we report are the average of 100 different channel realizations. The channel was normalized so
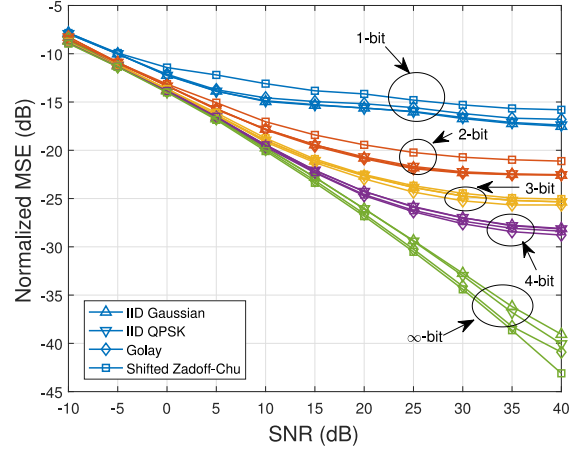

Fig. 4. EM-GM-VAMP recovery performance versus SNR for training matrix $\mathbf{T}$ constructed from IID Gaussian entries, IID QPSK entries, Golay complementary sequences, and shifted-ZC sequences. Here, $N_t = N_r = 64$, $L = 16$, $N_{cl} = 4$, and $N_p = 2048$.

that $\mathbb{E}\left[\sum_\ell \|\mathbf{H}[\ell]\|_F^2\right] = \mathbb{E}\left[\|\mathbf{x}\|^2\right] = N_t N_r$. We define

$$\mathrm{SNR} \triangleq \frac{\mathbb{E}[\|\mathbf{z}\|^2]}{\mathbb{E}[\|\mathbf{w}\|^2]} = \frac{\mathbb{E}[\mathrm{tr}\{\mathbf{A}\mathbf{x}\mathbf{x}^*\mathbf{A}^*\}]}{N_r N_p \sigma_w^2} = \frac{P_t L N_p N_r}{L N_r N_p \sigma_w^2} = \frac{P_t}{\sigma_w^2}.$$

A. Choice of Training Matrix

We first investigate the performance of various choices of training matrix $\mathbf{T}$. In addition to the circularly-shifted-ZC design proposed in Section IV-D, we try $\mathbf{T}$ constructed from Golay complementary sequences, IID random Gaussian entries, and IID random QPSK entries. The channel estimation NMSEs of EM-GM-VAMP with these $\mathbf{T}$ are compared in Fig. 4, where

$$\mathrm{NMSE}(\widetilde{\mathbf{x}}) \triangleq \mathbb{E}\left[\frac{\|\widetilde{\mathbf{x}} - \mathbf{x}\|^2}{\|\mathbf{x}\|^2}\right], \tag{47}$$

for $\widetilde{\mathbf{x}}$ normalized according to (46). As seen in the figure, the NMSEs of the four training designs are very close. However, the Golay sequence, random QPSK, and ZC sequences are preferred because of their constant modulus property, which leads to low PAPR. Furthermore, the ZC-based design allows efficient implementation of the matrix-vector multiplications in VAMP, as discussed in Section IV-D.

The runtime of the matrix-vector multiplications $\mathbf{A}\widehat{\mathbf{x}}$ and $\mathbf{A}^*\widehat{\mathbf{s}}$ and their fast implementations (using MATLAB R2016a on a standard desktop computer) are shown in Fig. 5. In the baseline implementation, $\mathbf{A}\widehat{\mathbf{x}}$ was computed via $\mathrm{vec}(\mathbf{B}_{N_r} \widehat{\mathbf{X}}\mathbf{C})$ to avoid generating and storing the high-dimensional matrix $\mathbf{A}$, and the computation of $\mathbf{A}^*\widehat{\mathbf{s}}$ was done similarly. The fast implementation was described in Section IV-D. As seen in the figure, the fast implementation consumes much less time than the baseline implementation. Much faster runtimes would be possible if the FFTs were implemented in hardware.

B. Algorithm Complexity

We now investigate the computational complexity of the various algorithms under test. Fig. 6 shows NMSE versus iteration for EM-GAMP, EM-VAMP, and QIHT. The figure shows the VAMP algorithms converging in ≈ 3 iterations, the GAMP
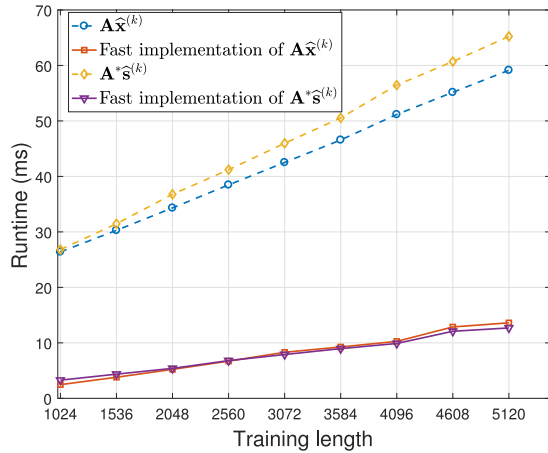
Fig. 5.   Runtime versus training length N_p to compute the matrix-vector multiplications $\mathbf{A}\widehat{\mathbf{x}}$ and $\mathbf{A}^*\widehat{\mathbf{s}}$ implemented via $\mathbf{A}\widehat{\mathbf{x}} = \text{vec}\left(\mathbf{B}_{N_r}\widehat{\mathbf{X}}\mathbf{C}\right)$ or via the fast method from Section IV-D. Here, $N_\text{t} = N_\text{r} = 64$ and $L = 16$, so that the dimensions of $\mathbf{A}$ are $(64N_\text{p}) \times 65536$.
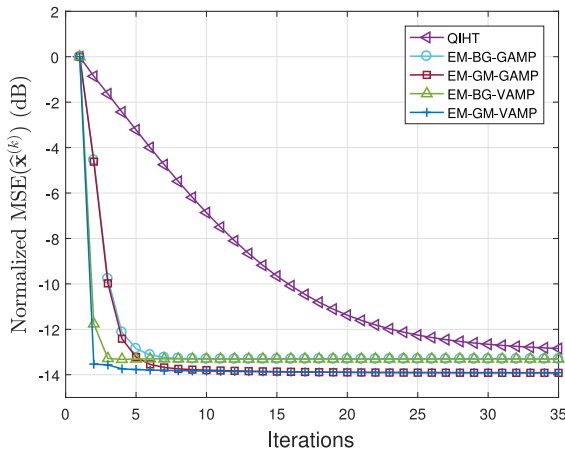

Fig. 6.   NMSE versus iteration for several algorithms. Here, SNR = 0 dB, the ADC had 4 bits, and the training was shifted ZC of length $N_\text{p} = 2048$.

algorithms converging in ≈ 8 iterations, and QIHT converging in ≈ 35 iterations.

To further investigate algorithm complexity, we plot NMSE versus runtime in Fig. 7. We controlled the runtime of the QIHT, EM-GAMP and EM-VAMP algorithms by varying the number of iterations. The other three (non-iterative) algorithms are each represented by a single point on the plot. The figure shows that EM-GM-VAMP gives the best NMSE-complexity trade-off for runtimes > 0.4 s, while EM-BG-VAMP gives the best trade-off for runtimes between 0.17 s and 0.4 s. The relatively long time required to complete the first iteration of EM-VAMP is likely due to the object-oriented MATLAB implementation[5] and would likely not be an issue in a dedicated implementation. Although QIHT can complete a few of its iterations before the first EM-BG-VAMP iteration, the corresponding estimates are probably not useful because their NMSE is so poor. The figure also shows that the NMSE-complexity frontier of EM-BG-GAMP is not too far away from that of EM-GM-VAMP, and

---

[5]MATLAB codes for EM-GAMP and EM-VAMP are available from http://sourceforge.net/projects/gampmatlab/.
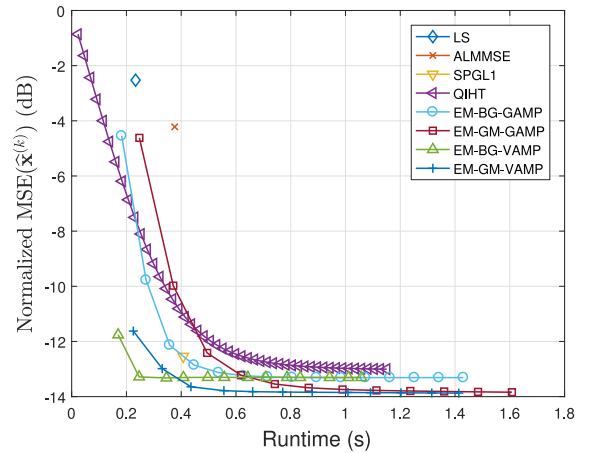

Fig. 7.   NMSE versus runtime for several algorithms, where the number of algorithm iterations was varied to obtain different runtimes. Here, SNR = 0 dB, the ADC had 4 bits, and the training was shifted ZC of length $N_\text{p} = 2048$.

TABLE II
ALGORITHMIC COMPLEXITY

Algorithm	Complexity Scaling
LS	$\mathcal{O}(N_\text{t}{}^3 L^3)$
ALMMSE	$\mathcal{O}(N_\text{t}{}^3 L^3)$
SPGL1	$\mathcal{O}\left(N_\text{t} N_\text{r} L \log\left(N_\text{t} N_\text{r} L\right)\right)$
QIHT	$\mathcal{O}\left(N_\text{t} N_\text{r} L \log\left(N_\text{t} N_\text{r} L\right)\right)$
EM-AMP	$\mathcal{O}\left(N_\text{t} N_\text{r} L \log\left(N_\text{t} N_\text{r} L\right)\right)$

that SPGL1 achieves one specific point on the EM-BG-GAMP frontier. Finally, the figure shows that the LS and ALMMSE algorithms are far from optimal.

Table II summarizes the complexity scalings of the algorithms under test. The complexities of the LS and ALMMSE approaches are dominated by the inversion of an $N_\text{p} \times N_\text{p}$ matrix, where N_p is an integer multiple of $N_\text{t}L$, and so their complexities scale as $\mathcal{O}(N_\text{t}{}^3 L^3)$. The complexities of the EM-AMP, SPGL1, and QIHT algorithms are dominated by the matrix-vector multiplications with $\mathbf{A}$ and $\mathbf{A}^*$. Matrix $\mathbf{A}$ is of size $N_x \times N_y$, where $N_x = N_\text{t} N_\text{r} L$ and $N_y = N_\text{r} N_\text{p}$. Since N_p is an integer multiple of $N_\text{t}L$, both N_x and N_y are $\mathcal{O}(N_\text{t} N_\text{r} L)$. And since we use an FFT-like algorithm to implement the matrix-vector multiplies with $\mathbf{A}$ and $\mathbf{A}^*$, the complexity order of EM-AMP, SPGL1, and QIHT scales as $\mathcal{O}\left(N_\text{t} N_\text{r} L \log(N_\text{t} N_\text{r} L)\right)$. In our simulations, N_t and N_r are of a similar size, and so we conclude that the complexity scaling of EM-AMP, SPGL1, and QIHT is much lower than that of LS and ALMMSE.

C. Effect of SNR, ADC Resolution, and Training Length

We now investigate estimation performance versus SNR, ADC resolution, and training length. Fig. 8 shows the NMSE of the VAMP algorithms versus SNR. The figure shows that, at low SNR (< 0 dB), the performance gap between 1-bit ADC and infinite-bit ADC is only 2 dB. But as the SNR increases, the gap between 1-bit and infinite-bit performance grows. Thus, higher resolution ADCs provide significant benefits only at higher SNRs. To reduce power consumption and cost, few-bit ADCs should be deployed when the SNR is low.
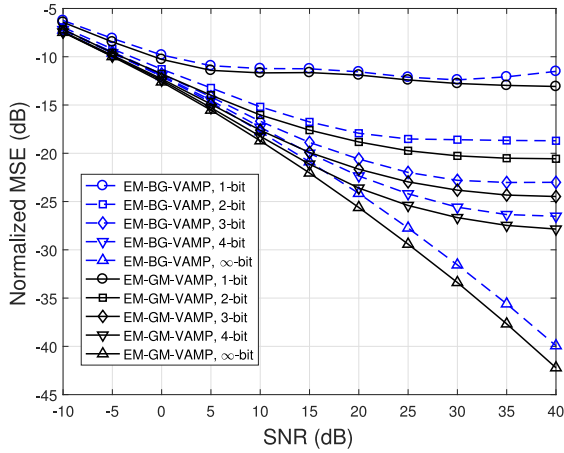
Fig. 8. NMSE versus SNR for the EM-VAMP algorithm under various ADC resolutions. Here, the training was shifted ZC with length $N_{\mathrm{p}} = 2048$.
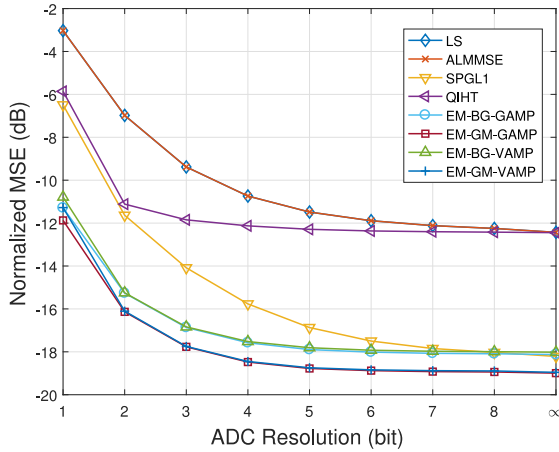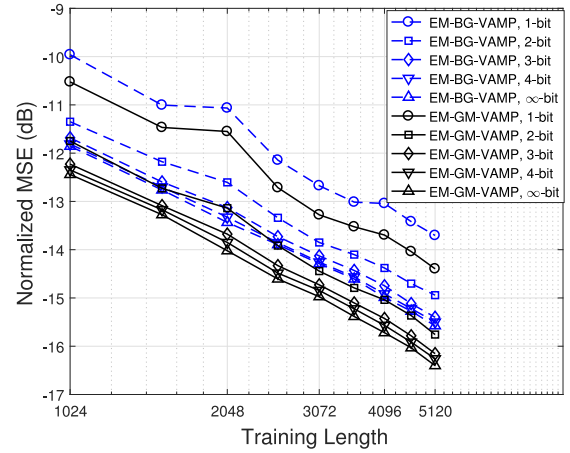

Fig. 10. NMSE versus training length for the EM-VAMP algorithm under various ADC resolutions. Here, the training was shifted ZC and SNR = 0 dB. It is observed that the estimation error exponentially decreases with the training length.

Fig. 9. NMSE versus ADC resolution for several algorithms. Here, SNR = 10 dB and the training was shifted ZC with length $N_{\mathrm{p}} = 2048$.

Fig. 9 shows NMSE versus ADC resolution for all algorithms under test when SNR = 10 dB. The figure shows that NMSE decreases with ADC resolution for all algorithms, but not in a uniform way. For example, SPGL1 and QIHT perform similarly with 1 and 2 bits of resolution, but SPGL1 benefits from higher ADC resolution while QIHT does not. In fact, SPGL1 performs as good as EM-BG-GAMP/VAMP for > 7 bits, but significantly worse with few bits, because it does not leverage the structure of the quantization error. The figure also shows that, when the EM-GAMP/VAMP are used, there is little benefit in increasing the ADC resolution above 4 bits at this SNR.

Fig. 10 shows the NMSE of the VAMP algorithms versus training length N_{p}. We note that, as the training length varies from 1024 to 5120, the sampling rate relative to Nyquist, i.e., $\frac{N_y}{N_x} = \frac{N_{\mathrm{p}} N_{\mathrm{r}}}{L N_{\mathrm{t}} N_{\mathrm{r}}} = \frac{N_{\mathrm{p}}}{1024}$, varies from 1 to 5. The figure shows that the NMSE decays polynomially with the training length. In particular, NMSE $\propto N_{\mathrm{p}}^{-\alpha}$ with $\alpha \approx 1/2$.

D. Mutual Information and Achievable Rate Bounds

In this section, we investigate the effect of channel-estimation accuracy on mutual information and achievable rate, which are important metrics for communication systems. Our methodology is inspired by that in [20].

We consider OFDM transmission with N_{b} subcarriers. For the kth subcarrier, denote the true frequency-domain MIMO channel by $\mathbf{G}_k \in \mathbb{C}^{N_{\mathrm{r}} \times N_{\mathrm{t}}}$, the transmitted signal by $\mathbf{s}_k$, and the variance-$\sigma_{w_{\mathrm{f}}}^2$ additive Gaussian noise by $\mathbf{w}_k$. Using the linearized model (35), the kth subcarrier output is

$$\mathbf{y}_k = (1 - \eta_b)\mathbf{G}_k\mathbf{s}_k + \widehat{\mathbf{w}}_k, \tag{48}$$

with effective noise $\widehat{\mathbf{w}}_k = (1 - \eta_b)\mathbf{w}_k + \mathbf{w}_{\mathrm{q},k}$. Using the Bussgang approximation (36), we can approximate its covariance (conditional on $\{\mathbf{G}_l\}_{l=1}^{N_{\mathrm{b}}}$) by

$$\mathbb{E}\left[\widehat{\mathbf{w}}_k\widehat{\mathbf{w}}_k^* \,|\, \{\mathbf{G}_l\}\right] \approx (1 - \eta_b)\Bigg[\sigma_{w_{\mathrm{f}}}^2 \mathbf{I}_{N_{\mathrm{r}}} \tag{49}$$

$$+ \eta_b \mathrm{diag}\left(\frac{1}{N_{\mathrm{b}}}\sum_{l=1}^{N_{\mathrm{b}}} \mathbf{G}_l \mathbf{R}_l \mathbf{G}_l^*\right)\Bigg] \triangleq \widehat{\boldsymbol{\Sigma}}_{\widehat{w}_k},$$

where $\mathbf{R}_k \triangleq \mathbb{E}[\mathbf{s}_k\mathbf{s}_k^*]$ and the averaging over N_{b} subcarriers occurs because quantization is performed in the time-domain.

Suppose that $\mathbf{s}_k$ is precoded based on the estimated frequency-domain channel $\widehat{\mathbf{G}}_k$, which is computed from the estimated time-domain channel $\{\widehat{\mathbf{H}}[\ell]\}_{\ell=0}^{L-1}$. In particular,

$$\mathbf{s}_k = \sum_{m=1}^{\min(N_{\mathrm{t}}, N_{\mathrm{r}})} \widehat{\mathbf{v}}_{km} \sqrt{p_{km}} d_{km}, \tag{50}$$

where $\widehat{\mathbf{v}}_{km}$ is the mth right-singular vector of $\widehat{\mathbf{G}}_k$, $p_{km} \geq 0$ are the powers allocated by the waterfilling algorithm, and d_{km} is a unit-variance message-bearing symbol.

If each stream is decoded independently and the effective noise $\widehat{\mathbf{w}}_k$ is treated as if it has the worst-case Gaussian distribution with covariance $\widehat{\boldsymbol{\Sigma}}_{\widehat{w}_k}$, then the mutual information between $\{\mathbf{s}_k\}_{k=1}^{N_{\mathrm{b}}}$ and $\{\mathbf{y}_k\}_{k=1}^{N_{\mathrm{b}}}$ conditioned on $\{\mathbf{G}_k\}_{k=1}^{N_{\mathrm{b}}}$ (in units of
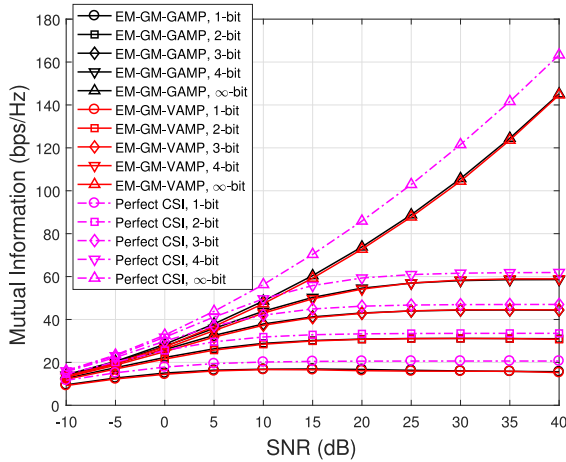
Fig. 11. Mutual information lower bound (51) versus SNR for EM-VAMP under various ADC resolutions. Here, the training was shifted ZC with length $N_p = 2048$.
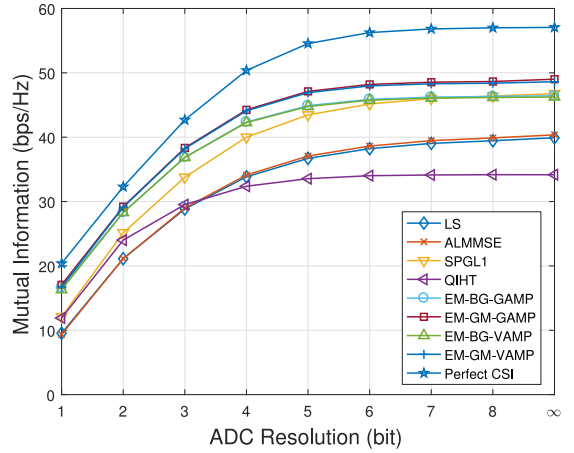

Fig. 12. Mutual information lower bound (51) versus ADC resolution for several algorithms. Here, SNR = 10 dB and the training was shifted ZC with length $N_p = 2048$.

bps/Hz) is lower bounded by [51]

$$
\underline{I}\left(\{\mathbf{s}_k\}_{k=1}^{N_b}, \{\mathbf{y}_k\}_{k=1}^{N_b} \,\middle|\, \{\mathbf{G}_k\}_{k=1}^{N_b}\right)
$$

$$
\triangleq \mathbb{E}\left\{ \frac{1}{N_b} \sum_{k=1}^{N_b} \sum_{m=1}^{\min(N_t, N_r)} \log_2 \left(1 \right.\right. \tag{51}
$$

$$
\left.\left. + \frac{(1-\eta_b)^2 |\widehat{\mathbf{u}}_{km}^* \mathbf{G}_k \widehat{\mathbf{v}}_{km}|^2 p_{km}}{\widehat{\mathbf{u}}_{km}^* \widehat{\boldsymbol{\Sigma}}_{\widehat{w}_k} \widehat{\mathbf{u}}_{km} + \sum_{n \neq m} (1-\eta_b)^2 |\widehat{\mathbf{u}}_{km}^* \mathbf{G}_k \widehat{\mathbf{v}}_{kn}|^2 p_{kn}} \right) \right\}
$$

where $\widehat{\mathbf{u}}_{km}$ is the mth left-singular vector of $\widehat{\mathbf{G}}_k$. The expectation in (51) is taken over $\{\mathbf{G}_k\}_{k=1}^{N_b}$ and the effective noise $\{\widehat{\mathbf{w}}_k\}_{k=1}^{N_b}$. (Because the channel estimates depend on $\mathbf{G}_k$ and $\widehat{\mathbf{w}}_k$, so do $\widehat{\mathbf{u}}_{km}$ and $\widehat{\mathbf{v}}_{km}$.) If the coherence time of the channel is N_{co} symbols, then a lower bound on the achievable rate (in bps/Hz) is given by [51]

$$
R \geq \frac{N_{co} - N_p}{N_{co}} \underline{I}\left(\{\mathbf{s}_k\}_{k=1}^{N_b}; \{\mathbf{y}_k\}_{k=1}^{N_b} \,\middle|\, \{\mathbf{G}_k\}_{k=1}^{N_b}\right), \tag{52}
$$

where $\frac{N_{co} - N_p}{N_{co}}$ represents loss due to training overhead.

Fig. 11 plots the mutual information lower bound (51) versus SNR under EM-GAMP/VAMP channel estimates and perfect CSI at various ADC resolutions, using Monte-Carlo to approximate the expectation. The figure shows that using EM-GAMP/VAMP channel estimates results in a relatively small loss in mutual information compared to perfect CSI. It also shows that, at low SNR, the mutual information loss of few-bit ADC relative to infinite-bit ADC is small.

Fig. 12 plots the mutual information lower bound (51) versus ADC resolution for all the channel-estimation algorithms under test at SNR = 10 dB. The figure shows that the EM-GM-GAMP/VAMP algorithms achieve the highest mutual information, with EM-BG-GAMP/VAMP close behind at low ADC resolutions. The figure also shows that the mutual information saturates when the ADC resolution is > 5 bits at this SNR.

Fig. 13 shows the achievable rate lower bound (52) versus training length N_p for the EM-VAMP algorithms. For this experiment, the channel coherence time was assumed to be $N_{co} = 10240$ symbols. The figure shows that, under these conditions,
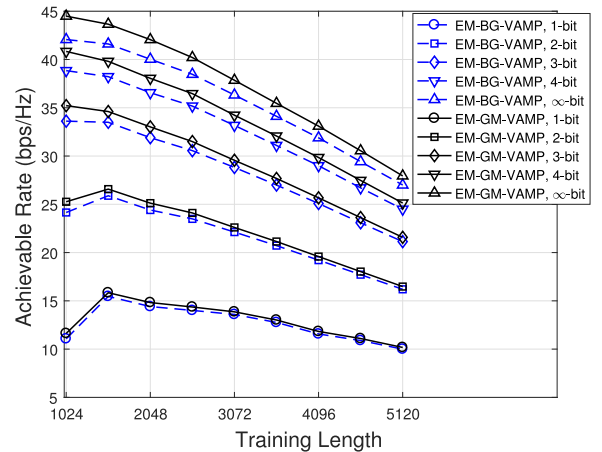
Fig. 13. Achievable rate lower bound (52) versus training length N_p for the EM-VAMP algorithm under various ADC resolutions. Here, SNR = 10 dB, the coherence time was $N_{co} = 10240$, and the training was shifted ZC.

the optimal training length is $N_p = 1536$ when the ADC has 1 or 2 bits of resolution. When the ADC resolution increases to 3 or 4 bits, a shorter training length (e.g., $N_p = 1024$) is preferred because the cost of each training symbol (in bps/Hz) is larger.

Fig. 14 shows the achievable rate lower bound (52) versus the training length N_p for all the channel estimation algorithms under test. For this experiment, the ADC has 2-bit resolution. The EM-GM-GAMP and EM-GM-VAMP algorithms provide highest achievable rate and their optimal training length is 1536. Note that, for the LS and ALMMSE algorithms, the optimal training length is 2048, which is longer than the other algorithms.

According to Figs. 13 and 14, the training overhead is around 10%–20% when the channel coherence length is 10240. The overhead percentage is thus comparable to that used in [47] for sub-6 GHz massive MIMO channel estimation with a flat fading channel [47]. Our channel, however, is frequency-selective fading and thus more difficult to estimate. In addition, a training duration of 1024–2048 symbols seems appropriate for mmWave broadband communication. Comparing to the 802.11ad standard, for example, if 8×8 UPA arrays are used with the standard DFT-based codebook, then 128 pilot frames, each consisting of
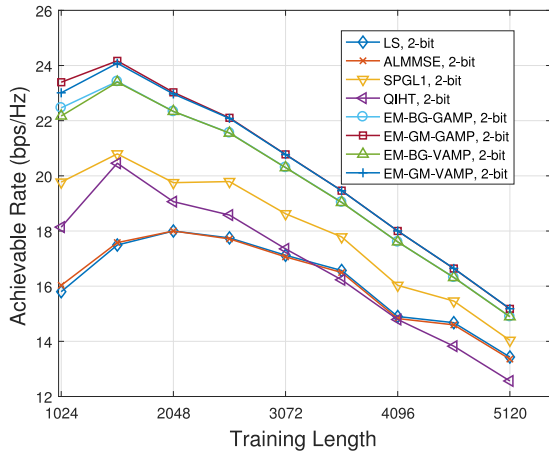
Fig. 14. Achievable rate lower bound (52) versus training length for several algorithms under 2-bit ADC. Here, SNR = 10 dB, the coherence time was $N_{co} = 10240$, and the training was shifted ZC.

26 bytes, would be transmitted [52]. The total training length in 802.11ad is thus much longer than what we propose.

VI. CONCLUSION

In this paper, we propose a methodology for estimation of broadband mmWave MIMO channels at receivers with few-bit ADCs. The broadband mmWave MIMO channel is sparse in both angle and delay domains, making it natural to apply compressed sensing techniques. We propose to use computationally efficient AMP algorithms (i.e., EM-GAMP and EM-VAMP) that accurately estimate the channel in the absence of prior information about its distribution (e.g., sparsity), and we enhance those methods with separate channel-norm estimation. We also design a training scheme, based on shifted ZC sequences, that leads to accurate and computationally efficient estimation with minimal transmitter PAPR. Finally, we report the results of an extensive simulation study that tests various algorithms, ADC precisions, training sequences, training lengths, SNRs, and runtime limits. Our simulations investigated channel-estimation MSE as well as mutual information and achievable rate bounds.

From the results of our study, we draw several conclusions. First, it is important to exploit the available joint angle-delay domain sparsity in channel estimation. Second, the ADC precision should be chosen based on the SNR; at low SNR, the use of few-bit ADCs results in very small loss (in MSE or achievable rate) compared to infinite-bit ADCs. Third, the training length should be chosen based on the ADC precision; at lower ADC precision, the achievable rate is maximized by a longer training sequence.

In this paper, the correlation among angle-delay channel coefficients was neglected. A possible direction to improve the channel estimation accuracy is to exploit this correlation.

REFERENCES

[1] J. Mo, P. Schniter, N. G. Prelcic, and R. W. Heath, Jr., "Channel estimation in millimeter wave MIMO systems with one-bit quantization," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 957–961.

[2] T. S. Rappaport, R. W. Heath, Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. Englewood Cliffs, NJ, USA: Prentice Hall, 2015.

[3] B. Murmann, "Energy limits in A/D converters," in *Proc. IEEE Faible Tension Faible Consommation*, Jun. 2013, pp. 1–4.

[4] B. Le, T. Rondeau, J. Reed, and C. Bostian, "Analog-to-digital converters," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 69–77, Nov. 2005.

[5] S. Jacobsson, G. Durisi, M. Coldrey, U. Gustavsson, and C. Studer, "One-bit massive MIMO: Channel estimation and high-order modulations," in *Proc. IEEE Intl. Conf. Commun. Workshop*, Jun. 2015, pp. 1304–1309.

[6] C. K. Wen, C. J. Wang, S. Jin, K. K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, May 2016.

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

[8] J. A. Nossek and M. T. Ivrlač, "Capacity and coding for quantized MIMO systems," in *Proc. Int. Conf. Wireless Commun. Mobile Comput.*, 2006, pp. 1387–1392.

[9] A. Mezghani and J. Nossek, "On ultra-wideband MIMO systems with 1-bit quantized outputs: Performance analysis and input optimization," in *Proc. IEEE Int. Symp. Inf. Theory*, 2007, pp. 1286–1289.

[10] A. Mezghani and J. Nossek, "Capacity lower bound of MIMO channels with output quantization and correlated noise," in *Proc. IEEE Int. Symp. Inf. Theory*, 2012, pp. 1–5.

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

[12] O. Dabeer and U. Madhow, "Channel estimation with low-precision analog-to-digital conversion," in *Proc. IEEE Int. Conf. Commun.*, 2010, pp. 1–6.

[13] G. Zeitler, G. Kramer, and A. Singer, "Bayesian parameter estimation using single-bit dithered quantization," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2713–2726, Jun. 2012.

[14] C. Risi, D. Persson, and E. G. Larsson, "Massive MIMO with 1-bit ADC," arXiv preprint arXiv:1404.7736, Apr. 30, 2014.

[15] S. Wang, Y. Li, and J. Wang, "Multiuser detection in massive spatial modulation MIMO with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2156–2168, Apr. 2015.

[16] M. T. Ivrlac and J. A. Nossek, "On MIMO channel estimation with single-bit signal-quantization," in *Proc. ITG Workshop Smart Antennas*, Feb. 2007, pp. 1–7.

[17] A. Mezghani, F. Antreich, and J. Nossek, "Multiple parameter estimation with quantized channel output," in *Proc. ITG Workshop Smart Antennas*, 2010, pp. 143–150.

[18] T. Rappaport, G. Maccartney, M. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.

[19] "5G channel model for bands up to 100 GHz (v2.3)," Aalto Univ., Espoo, Finland, AT&T, BUPT, CMCC, Ericsson, Huawei, Intel, KT Corporation, Nokia, NTT DOCOMO, New York Univ., New York, NY, USA, Qualcomm, Samsung, Univ. Bristol, Bristol, U.K., and Univ. Southern California, Los Angeles, CA, USA, Tech. Rep., 2016. [Online]. Available: http://www.5gworkshops.com/5gcm.html

[20] P. Schniter and A. Sayeed, "Channel estimation and precoder design for millimeter-wave communications: The sparse way," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 273–277.

[21] H. Wang, T. Liu, C. K. Wen, and S. Jin, "Optimal data detection for OFDM system with low-resolution quantization," in *Proc. IEEE Int. Conf. Commun. Syst.*, Dec. 2016, pp. 1–6.

[22] T. C. Zhang, C. K. Wen, S. Jin, and T. Jiang, "Mixed-ADC massive MIMO detectors: Performance analysis and design optimization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7738–7752, Nov. 2016.

[23] C. Stöckle, J. Munir, A. Mezghani, and J. A. Nossek, "Channel estimation in massive MIMO systems using 1-bit quantization," in *Proc. IEEE Workshop Signal. Process. Adv. Wireless Commun.*, Jul. 2016, pp. 1–6.

[24] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2011, pp. 2168–2172.

[25] S. Rangan, P. Schniter, and A. Fletcher, "Vector approximate message passing," arXiv preprint arXiv:1610.03082, Oct. 10, 2016.

[26] P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, 2016, pp. 1525–1529.

[27] J. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.

[28] A. K. Fletcher and P. Schniter, "Learning and free energies for vector approximate message passing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 4247–4251.

[29] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, 1960.

[30] X.-P. Chen, K. Wu, L. Han, and F. He, "Low-cost high gain planar antenna array for 60-GHz band applications," *IEEE Trans Antennas Propag.*, vol. 58, no. 6, pp. 2126–2129, Jun. 2010.

[31] F. Guidi, A. Guerra, and D. Dardari, "Personal mobile radars with millimeter-wave massive arrays for indoor mapping," *IEEE Trans. Mobile Comput.*, vol. 15, no. 6, pp. 1471–1484, Jun. 2016.

[32] W. Hong, K.-H. Baek, Y. Lee, Y. Kim, and S.-T. Ko, "Study and prototyping of practically large-scale mmWave antenna systems for 5G cellular devices," *IEEE Commun. Mag.*, vol. 52, no. 9, pp. 63–69, Sep. 2014.

[33] A. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[34] J. Brady and A. Sayeed, "Beamspace MU-MIMO for high-density gigabit small cell access at millimeter-wave frequencies," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun.*, Jun. 2014, pp. 80–84.

[35] L. Jacques, J. Laska, P. Boufounos, and R. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2082–2102, Apr. 2013.

[36] L. Jacques, K. Degraux, and C. De Vleeschouwer, "Quantized iterative hard thresholding: Bridging 1-bit and high-resolution quantized compressed sensing," arXiv preprint arXiv:1305.1786, May 8, 2013.

[37] A. Zymnis, S. Boyd, and E. Candes, "Compressed sensing with quantized measurements," *IEEE Signal Process. Lett.*, vol. 17, no. 2, pp. 149–152, Feb. 2010.

[38] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 482–494, Jan. 2013.

[39] U. Kamilov, V. Goyal, and S. Rangan, "Message-passing de-quantization with applications to compressed sensing," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6270–6281, Dec. 2012.

[40] A. Mezghani and J. Nossek, "Efficient reconstruction of sparse vectors from quantized observations," in *Proc. ITG Workshop Smart Antennas*, Mar. 2012, pp. 193–200.

[41] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Inf. Inference*, vol. 2, no. 2, pp. 115–144, 2013.

[42] U. S. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2969–2985, May 2014.

[43] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. Buckinghamshire, U.K.: Univ. Press Group Limited, 2006.

[44] S. Rangan, P. Schniter, and A. Fletcher, "On the convergence of approximate message passing with arbitrary matrices," in *Proc. IEEE Int. Symp. Info. Theory.*, Jun. 2014, pp. 236–240.

[45] D. Chu, "Polyphase codes with good periodic correlation properties," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 531–532, Jul. 1972.

[46] *Physical Channels and Modulation*, 3GPP TS 36.211 V12.7.0 Std., 2015.

[47] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, Aug. 2017.

[48] C. Mollen, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 87–100, Jan. 2017.

[49] K. Roth and J. A. Nossek, "Achievable rate and energy efficiency of hybrid and digital beamforming receivers with low resolution ADC," arXiv preprint arXiv:1610.02909, Oct. 10, 2016.

[50] E. van den Berg and M. P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. Sci. Comput.*, vol. 31, no. 2, pp. 890–912, 2009.

[51] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge Univ. Press, 2005.

[52] T. Nitsche, C. Cordeiro, A. B. Flores, E. W. Knightly, E. Perahia, and J. C. Widmer, "IEEE 802.11ad: Directional 60 GHz communication for multigigabit-per-second wi-fi [invited paper]," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 132–141, Dec. 2014.
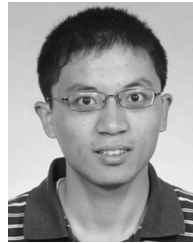
Jianhua Mo (S'12–M'17) received the B.S. and M.S. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2010 and 2013, respectively, and the Ph.D. degree from the University of Texas at Austin, Austin, TX, USA, in 2017, all in electrical engineering. He is currently a Senior Engineer with Samsung Research America, Dallas, TX, USA. His areas of interests included the physical layer security, MIMO communications with low-resolution ADCs. He is currently working on the mmWave wireless backhaul transmission. He was a co-recipient of the Heinrich Hertz Award for best communications letter in 2013. He was recognized as an exemplary reviewer of the IEEE WIRELESS COMMUNICATIONS LETTERS in 2012 and the IEEE COMMUNICATIONS LETTERS in 2015.

Philip Schniter (S'92–M'93–SM'05–F'14) received the B.S. and M.S. degrees in electrical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992 and 1993, respectively, and the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, USA, in 2000. From 1993 to 1996, he was employed with Tektronix Inc., Beaverton, OR, USA, as a Systems Engineer. After receiving the Ph.D. degree, he joined the Department of Electrical and Computer Engineering, Ohio State University, Columbus, OH, USA, where he is currently a Professor. In 2008–2009, he was a Visiting Professor with Eurecom, Sophia Antipolis, France, and with Supelec, Gif-sur-Yvette, France. In 2016–2017, he was a Visiting Professor with Duke University, Durham, NC, USA. His areas of interests currently include signal processing, wireless communications, and machine learning. In 2002, he was the recipient of the NSF CAREER Award, in 2016 the IEEE Signal Processing Society Best Paper Award, and in 2017 the Qualcomm Faculty Award. He currently serves on the IEEE Sensor Array and Multichannel Technical Committee and the IEEE Computational Imaging Special Interest Group.

Robert W. Heath, Jr. (S'96–M'01–SM'06–F'11) received the B.S. and M.S. degrees from the University of Virginia, Charlottesville, VA, USA, in 1996 and 1997, respectively, and the Ph.D. degree from Stanford University, Stanford, CA, USA, in 2002, all in electrical engineering. From 1998 to 2001, he was a Senior Member of the Technical Staff then a Senior Consultant with Iospan Wireless Inc., San Jose, CA, USA, where he worked on the design and implementation of the physical and link layers of the first commercial MIMO-OFDM communication system. Since January 2002, he has been with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA, where he is a Cullen Trust for Higher Education Endowed Professor, and is a Member of the Wireless Networking and Communications Group. He is also the President and CEO of MIMO Wireless Inc. He has authored "Introduction to Wireless Digital Communication" (Prentice Hall, 2017) and "Digital Wireless Communication: Physical Layer Exploration Lab Using the NI USRP" (National Technology and Science Press, 2012), and co-authored "Millimeter Wave Wireless Communications" (Prentice Hall, 2014). He was a Distinguished Lecturer in the IEEE Signal Processing Society and is an ISI Highly Cited Researcher. He is also an elected member of the Board of Governors for the IEEE Signal Processing Society, a licensed Amateur Radio Operator, a Private Pilot, and a registered Professional Engineer in Texas. He has been a co-author of a number of award winning conference and journal papers including the 2010 and 2013 *EURASIP Journal on Wireless Communications and Networking* best paper awards, the 2012 Signal Processing Magazine best paper award, a 2013 Signal Processing Society best paper award, 2014 *EURASIP Journal on Advances in Signal Processing* best paper award, the 2014 *Journal of Communications and Networks* best paper award, the 2016 IEEE Communications Society Fred W. Ellersick Prize, the 2016 IEEE Communications and Information Theory Societies Joint Paper Award, and the 2017 Marconi Prize Paper Award. He was the recipient of the 2017 EURASIP Technical Achievement award.