SPADE: Sparsity-Adaptive Equalization for MMwave Massive MU-MIMO

Conference Paper

Author(s):
Mirfarshbafan, Seyedhadi; Studer, Christoph

Publication date:
2021

Permanent link:
https://doi.org/10.3929/ethz-b-000504110

Rights / license:
In Copyright - Non-Commercial Use Permitted

Originally published in:
https://doi.org/10.1109/SSP49050.2021.9513825
We propose SParsity-ADaptive Equalization (SPADE), a novel approach to reduce the effective number of multiplications in sparse inner products by adaptively skipping multiplications that have little to no effect on the result. We apply SPADE to beamspace linear minimum mean square error (LMMSE) spatial equalization in all-digital millimeter-wave (mmWave) massive multiuser multiple-input multiple-output (MU-MIMO) systems. We propose a SPADE-based architecture that mutes insignificant multiplications to offer power savings. We use simulation results with line-of-sight (LoS) and non-LoS mmWave channel models to demonstrate that SPADE-LMMSE performs on par with state-of-the-art beamspace equalizers in terms of bit error-rate, while requiring significantly lower preprocessing complexity.

1. INTRODUCTION

Millimeter-wave (mmWave) communication [1, 2] and massive multiuser (MU) multiple-input multiple-output (MIMO) [3] are two core technologies of fifth-generation (5G) and beyond 5G wireless systems. While mmWave communication provides access to large portions of unused bandwidth, it suffers from high propagation losses [4]. MU-MIMO [3] is able to (i) compensate for the high propagation losses via fine-grained beamforming and (ii) support simultaneous communication with multiple user equipments (UEs) in the same frequency band. However, the large number of basestation (BS) antennas combined with the high baseband sampling rates create new challenges for mmWave massive MU-MIMO hardware design.

While hybrid digital-analog architectures [5, 6] can reduce the hardware complexity, all-digital architectures with low-resolution data converters [7–9] and low-resolution baseband processing [10] achieve higher spectral efficiency, provide more flexibility, and simplify radio-frequency (RF) circuitry and baseband processing [11, 12]. However, to reduce the complexity and power consumption of baseband processing in all-digital BS designs, novel algorithms and architectures are necessary. Fortunately, mmWave channels typically consist of only a few dominant propagation paths [1, 4], which can be exploited to simplify BS design [13–18]. Concretely, by taking a spatial discrete Fourier transform (DFT) over the antenna array, which converts the antenna domain into beamspace domain, one can drastically simplify some of the most complex baseband processing tasks, including channel estimation [18] and spatial equalization [19–21]. Existing beamspace equalizers either suffer from a notable performance degradation compared to antenna domain spatial equalizers [20] or result in high preprocessing complexity [21].

1.1. Contributions

We propose SParsity-ADaptive Equalization (SPADE), a novel scheme to reduce the effective number of multiplications in sparse inner products. SPADE compares the entries of the two vectors and adaptively skips scalar multiplications which have negligible impact on the final result. When applied to beamspace linear minimum mean square error (LMMSE) equalization, SPADE offers significant reductions in the number of multiplications, without increasing the preprocessing complexity of LMMSE. We propose an architecture for SPADE-LMMSE that adaptively mutes multiplications to reduce power consumption. Furthermore, we use simulations with line-of-sight (LoS) and non-LoS mmWave channels to demonstrate that SPADE performs on par with the state-of-the-art beamspace equalization algorithms.

1.2. Notation

Boldface lowercase and uppercase letters represent column vectors and matrices, respectively. For a matrix \(\mathbf{A}\), the transpose and Hermitian transpose are \(\mathbf{A}^\mathsf{T}\) and \(\mathbf{A}^\mathsf{H}\), respectively, the \(k\)th column is denoted by \(\mathbf{a}_k\), and the Frobenius norm is \(\|\mathbf{A}\|_F\).

For a vector \(\mathbf{a}\), the \(k\)th entry is denoted by \(a_k\), the real and imaginary parts are denoted by \(a_\mathbb{R}\) and \(a_\mathbb{I}\), respectively, and the \(\ell_\infty\)-norm and \(\ell_\infty\)-norm are defined as \(\|\mathbf{a}\|_\infty = \max_k |a_k|\) and \(\|\mathbf{a}\|_\infty = \max\{\|a_\mathbb{R}\|, \|a_\mathbb{I}\|\}\), respectively [22]. The \(N \times N\) identity and the unitary \(N \times N\) discrete Fourier transform (DFT) matrices are denoted by \(\mathbf{I}_N\) and \(\mathbf{F}_N\), respectively.

2. BACKGROUND

2.1. Antenna-Domain and Beamspace System Models

We focus on an all-digital mmWave massive MU-MIMO uplink system as depicted in Figure 1. We assume frequency-flat
channels in which \( U \) single-antenna user equipments (UEs) transmit information simultaneously to a \( B \)-antenna base-station (BS) in the same frequency band. We model the antenna domain received signal vector \( y \in \mathbb{C}^B \) at the BS as \( y = \hat{H}s + n \), where \( \hat{H} \in \mathbb{C}^{B \times U} \) is the uplink channel matrix, \( s \in \mathbb{S}^U \) is the vector of the UEs’ data symbols taken from a discrete constellation set \( \mathbb{S} \) (e.g., 16-QAM) with \( E_s = \mathbb{E} |s|^2 \), \( u = 1, \ldots, U \). The noise vector \( n \in \mathbb{C}^B \) has i.i.d. circularly-symmetric complex Gaussian entries with variance \( \sigma_n^2 \). The average receive signal-to-noise ratio (SNR) at the BS is defined as \( SNR \triangleq E_s \| \hat{H} \|_F^2 / (BN_0) \).

In what follows, we consider mmWave propagation conditions with a \( B \)-antenna uniform linear array (ULA). To obtain the beamspace input-output relation, one applies a spatial DFT [23] to the received antenna-domain vector \( y \) as follows:

\[
y = F_B \tilde{y} = \hat{H}s + n. \tag{1}
\]

Here, \( y = F_B \tilde{y} \) is the beamspace receive vector, \( \hat{H} = F_B \tilde{H} \) is the (typically sparse) beamspace channel matrix, and \( n \) is the beamspace noise vector with the same statistics as the antenna-domain noise vector \( n \). Since the beamspace DFT must be computed for every receive vector \( \tilde{y} \), a practical system would rely on streaming fast Fourier transforms (FFTs) [23].

To see why the beamspace transform sparsifies mmWave channel matrices, consider the following plane-wave model for the antenna-domain channel vector \( \hat{h}_u \) of the \( u \)th UE [24]:

\[
\hat{h}_u = \sum_{\ell=0}^{L-1} \alpha_{\ell} \mathbf{a}(\phi_{\ell}). \tag{2}
\]

Here, \( L \) stands for the number of propagation paths, \( \alpha_{\ell} \in \mathbb{C} \) is the channel gain of the \( \ell \)th propagation path, and

\[
\mathbf{a}(\phi_{\ell}) = \left[ 1, e^{j \phi_{\ell}}, e^{j 2 \phi_{\ell}}, \ldots, e^{j (B-1) \phi_{\ell}} \right], \tag{3}
\]

where the spatial frequency \( \phi_{\ell} \) is determined by the \( \ell \)th path’s incident angle to the BS. Since \( L \) is small for line-of-sight (LoS) mmWave channels, taking a DFT \( \hat{h}_u = F_B \hat{h}_u \) reveals the sparse structure of such channel vectors, i.e., most of the vector’s energy is concentrated on the entries of \( \hat{h}_u \) associated with the spatial frequencies \( \phi_{\ell} \).

2.2. Beamspace LMMSE Equalization

Linear data detection consists of two phases: (i) \textit{preprocessing}, which is performed only once per channel coherence interval and produces a \( U \times B \) equalization matrix \( W^H \) and (ii) \textit{spatial equalization}, which is performed at baseband sampling rate (for each received signal vector \( y \)) in order to obtain estimates of the transmitted symbol vectors according to \( \hat{s} = W^H y \). In what follows, we focus on beamspace LMMSE equalization for which the equalization matrix is given by

\[
W^H = \left( H^H H + \frac{N_0}{C_s} I_U \right)^{-1} H^H. \quad \tag{4}
\]

In order to support high-bandwidth communication at mmWave frequencies, the spatial equalization stage must be carried out at extremely high baseband sampling rates. Hence, to keep power consumption within reasonable bounds, efficient methods to calculate \( \hat{s} = W^H y \) must be deployed in practice.

2.3. Existing Sparsity-Exploiting Spatial Equalizers

In recent years, a number of sparsity-exploiting beamspace equalization methods have been proposed [19–21]. All of these methods exploit the fact that for sparse beamspace channel matrices \( \hat{H} \), the associated LMMSE equalization matrices \( W^H \) tend to be sparse as well. This observation enables design of beamspace equalization algorithms that produce sparse equalization matrices \( W^H \) with a given density coefficient \( \delta \triangleq \| W \|_0 / (BN_0) \), where \( \| W \|_0 \) is the number of nonzero entries of \( W \). Such sparsity-exploiting spatial equalizers reduce the number of multiplications required when calculating \( \hat{s} = W^H y \), which reduces power consumption and/or implementation complexity. Among such methods, the entry-wise orthogonal matching pursuit (EOMP) proposed in [21] achieves the highest sparsity (lowest \( \delta \)) and hence highest complexity reduction during spatial equalization. EOMP and related algorithms, however, considerably increase the complexity of preprocessing, resulting in inefficient circuitry. We next propose SPADE-LMMSE, a novel sparsity-adaptive spatial equalization method that competes with EOMP in terms of complexity reduction during spatial equalization while directly using the conventional LMMSE equalization matrix (4) which means that the preprocessing complexity does not increase.

3. SPADE: SPARSITY-ADAPTIVE EQUALIZATION

Consider the inner product of two \( B \)-dimensional real-valued vectors \( \langle w, y \rangle = \sum_{b=1}^{B} w_b y_b \). Intuitively, if \( B \) is large, then we can skip partial products \( w_b y_b \) of small magnitude, without incurring a large relative error in the result, assuming that the exact result is bounded away from zero. However, we cannot eliminate partial products based on their magnitude, as this requires actually performing the multiplication. Therefore, we propose to set thresholds for the absolute values of the two operands \( w_b \) and \( y_b \), and skip (or mute) multiplications if the absolute values of both operands are below their respective thresholds. The same approach can be extended to complex-valued case, noting that each complex-valued inner product can be decomposed into four real-valued inner products. Since for mmWave channels, the rows \( w_u^H, u = 1, \ldots, U \) of beamspace
LMMSE matrices $W^H$ in (4) and the receive vectors $y$ exhibit sparsity, a large number of partial products when computing $s_u = w_u^H y$, $u = 1, \ldots, U$, will be small; hence, a large number of multiplications can be skipped—this is the general idea behind SPADE. In order to simplify hardware implementation of SPADE, we propose to use two fixed thresholds $\tau_y \in \mathbb{R}$ and $\tau_w \in \mathbb{R}$, for the real and imaginary parts of all entries of the receive vector $y$ and the equalization matrix $W^H$, respectively.

### 3.1. Setting the SPADE Thresholds

The thresholds $\tau_y$ and $\tau_w$ are used to trade arithmetic precision for reduction in the effective number of multiplications. Setting these thresholds close to zero will result in high precision, but will increase the number of active multiplications. In contrast, setting these thresholds to large values will result in precision loss, but will lower the number of active multiplications. To improve the effectiveness of SPADE, we propose the following techniques:

1) Since we use a single threshold $\tau_y$ for all the entries of $y$, and the statistics of $y$ change dynamically, we propose to determine a fixed threshold $\tau'_y$ and then set $\tau_y = \|y\|_{\infty} \tau'_y$ in order to incorporate fluctuations of $y$ into $\tau_y$.

2) Since we use a single threshold $\tau_w$ for all the entries of $W^H$, we scale the rows $w_u^H$, $u = 1, \ldots, U$, so that their $\ell_{\infty}$-norms are equal. Such a row-scaling scheme was proposed in [10] to reduce the dynamic range of the entries of equalization matrices. For SPADE, we scale the rows of the LMMSE equalization matrix according to $V = \text{diag}(\alpha)W^H$ such that the rows $V^H_u$, $u = 1, \ldots, U$, of the scaled matrix $V^H$ satisfy $\|v_u\|_{\infty} < 1$. This can be accomplished by $\alpha_u = 1/(\|w_u^H\|_{\infty} + \varepsilon)$, where $\varepsilon > 0$ is a small constant that ensures that $\|v_u\|_{\infty}$ is just below one. With this approach, estimates of the transmit vector are computed as $\hat{s} = \text{diag}(\alpha)^{-1}V^H y$, which corresponds to post-multiplying the $u$th entry of $V^H y$ by $1/\alpha_u$ for $u = 1, \ldots, U$. The threshold $\tau_w$ is applied to the entries of the scaled matrix $V^H$.

### 3.2. SPADE-Based Architecture

Figure 2 shows a high-level architecture of a fully-unrolled beamspace LMMSE equalizer employing SPADE to adaptively mute multipliers. We emphasize that in high-bandwidth mmWave systems with multi-GHz baseband sampling rates, fully-unrolled architectures become a natural choice to deliver the desired throughput while minimizing data buffering and control overhead [23]. In Figure 2, the LMMSE preprocessing block receives beamspace channel estimates and computes $V^H$ once per channel coherence interval. This block also performs the comparison of real and imaginary entries of $V^H$ with $\tau_w$ and provides the comparison bits to the SPADE-based equalizer block. The FFT block takes the antenna domain received vectors $\hat{y}$ and produces the beamspace domain vectors $\hat{y}$, along with the comparison bits for each real and imaginary part of entries of $\hat{y}$ with $\tau_y$. The fully-unrolled matrix-vector multiplication engine comprises $U$ inner product engines, each consisting of $B$ processing elements (PEs). Each PE is a complex-valued multiplier with four real-valued mutable multipliers (labeled as “MM”), whose internal architecture is depicted in Figure 2. At the beginning of each channel coherence interval, the weight registers of the PEs are loaded with the entries of $V^H$ and the comparison bits provided by the LMMSE preprocessing block. Then, for each beamspace receive vector coming from the FFT block, each MM conditionally disables the registers before and after the multiplier if the comparison bits indicate that both operands have absolute values smaller than their thresholds. Consequently, such a matrix-vector multiplication engine adaptively saves power by nulling the switching activity of unused multipliers—this will reduce the dynamic power consumption. For correct functionality, the output of each muted multiplier must be set to zero, which is implemented by the AND gate at the output of each MM as shown in Figure 2. The number of muted multipliers depends on the sparsity of $y$ and $V^H_u$, $u = 1, \ldots, U$, which is determined by the channel conditions. For example, as shown in Section 4, in a system with $B = 128$ BS antennas and $U = 16$ UEs with LoS channel conditions, it is possible to mute 80% of the multiplications while incurring no more than 0.1 dB SNR loss at an uncoded bit-error-rate (BER) of 1%.

We reiterate that LMMSE preprocessing is carried out only once per channel coherence interval and the spatial FFT can be implemented efficiently using the fully-unrolled multiplierless architecture proposed in [23]. However, the matrix-vector multiplication engine must operate constantly and at baseband.
We consider a mmWave massive MU-MIMO basestation with 4.1. Comparison with EOMP (which coincides with the density coefficient but is omitted due to space constraints. The values of the threshold pair an even lower uncoded BER of 0.1% BER of

\[ \text{uncoded BER of} \]

up to negligible SNR loss (compared to antenna domain LMMSE), multipliers for LoS channels than for non-LoS channels. For a also see that SPADE-LMMSE and EOMP require fewer active multipliers. The same threshold parameters in the non-LoS setting results in an SNR gap of less than 0.15 dB while requiring only 45% active multipliers. Another key advantage is that SPADE simply uses conventional LMMSE preprocessing whereas the preprocessing complexity of EOMP is considerably higher. Using the complexity expressions provided in [21], the preprocessing complexity of EOMP—measured in terms of the number of real-valued multiplications—is about 15× higher than that of conventional LMMSE for a \( B = 128 \) BS antenna and \( U = 16 \) UE system with density factor \( \delta = 20\% \).

5. CONCLUSIONS

We have proposed SParsity-ADaptive Equalization (SPADE), a novel spatial equalization approach that adaptively reduces the number of multiplications based upon the instantaneous channel conditions. SPADE offers the following advantages: (i) the preprocessing complexity is the same as the conventional LMMSE equalization, (ii) the performance degradation with respect to antenna-domain equalization is negligible for suitably chosen thresholds, and (iii) the method adaptively reduces complexity and lowers power based on the instantaneous channel conditions. For LoS and non-LoS mmWave massive MU-MIMO channel models, we have demonstrated that SPADE performs on par with EOMP [21], but requires significantly lower preprocessing complexity. We emphasize that SPADE is not only suitable for spatial equalization in mmWave massive MU-MIMO systems, but finds potential use in many other applications that carry out approximate sparse matrix-vector products. A hardware implementation of the proposed SPADE-LMMSE is ongoing work.

Fig. 3. Trade-off between SNR operating point at 1% uncoded BER and multiplier activity rate.
6. REFERENCES


