Journal of Signal Processing Systems manuscript No.
(will be inserted by the editor)

Implicit vs. Explicit Approximate Matrix Inversion for Wideband Massive MU-MIMO Data Detection

Michael Wu · Bei Yin · Kaipeng Li · Chris Dick · Joseph R. Cavallaro · Christoph Studer

December 2017

Abstract Massive multi-user (MU) MIMO wireless technology promises improved spectral efficiency compared to that of traditional cellular systems. While data-detection algorithms that rely on linear equalization achieve near-optimal error-rate performance for massive MU-MIMO systems, they require the solution to large linear systems at high throughput and low latency, which results in excessively high receiver complexity. In this paper, we investigate a variety of exact and approximate equalization schemes that solve the system of linear equations either *explicitly* (requiring the computation of a matrix inverse) or *implicitly* (by directly computing the solution vector). We analyze the associated performance/complexity trade-offs, and we show that for small base-station (BS)-to-user-antenna ratios, exact and implicit data detection using the Cholesky decomposition achieves near-optimal performance at low complexity. In contrast, implicit data detection using approximate equalization methods results in the best trade-off for large BS-to-user-antenna ratios. By combining the advantages of exact, approximate, implicit, and explicit matrix inversion, we develop a new frequency-adaptive equalizer (FADE), which outperforms existing data-detection methods in terms of performance *and* complexity for wideband massive MU-MIMO systems.

M. Wu, B. Yin, K. Li, and J. Cavallaro
Department of ECE, Rice University, Houston, TX
E-mail: {mbw2, by2, kl33, cavallar}@rice.edu

M. Wu and C. Dick
Xilinx Inc., San Jose, CA
E-mail: chris.dick@xilinx.com

C. Studer
School of ECE, Cornell University, Ithaca, NY
E-mail: studer@cornell.edu

Keywords Equalization · linear data detection · massive multi-user MIMO · matrix inversion · Neumann series expansion · SC-FDMA · OFDM

1 Introduction

Massive multi-user (MU) multiple-input multiple-output (MIMO) will be a core technology for fifth-generation (5G) wireless systems as it promises significant improvements in terms of the spectral efficiency compared to traditional, small-scale cellular MIMO technology [1–5]. The key idea of massive MU-MIMO is to deploy hundreds of antennas at the base station (BS) and to serve tens of single-antenna users concurrently and in the same frequency resource. This technology not only promises significant improvements in terms of the spectral efficiency compared to traditional, small-scale MIMO through fine-grained beamforming, but also enables simple and low-complexity data detection methods to achieve near-optimal error-rate performance [1]. However, for wideband massive MU-MIMO wireless systems with a large number of subcarriers, such as long-term evolution (LTE)-based systems with thousands of subcarriers, even some of the least expensive linear data-detection algorithms, e.g., methods relying on linear minimum mean square error (MMSE) equalization, require excessive hardware complexity and power consumption (see [6] for a detailed discussion).

In order to address the complexity and power consumption issue of linear data detection in wideband massive MU-MIMO systems, a variety of approximate matrix inversion methods have been proposed in recent years [1, 6–11]. These methods require, in general, lower computational complexity than exact, linear methods or non-linear algorithms, and entail only a small error-

rate performance loss in massive MU-MIMO systems having a large BS-to-user-antenna ratio. For systems in which the BS-to-user-antenna ratio is below two, however, approximate methods result in a strong error floor—often too high to enable reliable communication at high data rates. Furthermore, corresponding high-throughput very-large scale integration (VLSI) designs for wideband massive MU-MIMO systems that use orthogonal frequency-division multiplexing (OFDM) or single-carrier frequency-division multiple access (SC-FDMA), such as the reference hardware designs in [6,12], still require large silicon area and excessively high power consumption. Hence, to successfully deploy massive MU-MIMO in practical wideband systems, new algorithm solutions that achieve near-optimal error-rate performance at low hardware complexity are necessary.

1.1 Contributions

In this paper, we propose a host of novel low-complexity, soft-output data detection methods for wideband massive MU-MIMO systems that further reduce the complexity compared to existing approximate methods in [1, 6–11]. Our contributions are summarized as follows:

- We propose *accelerated* explicit matrix inversion methods building on the Neumann series expansion.
- We propose corresponding *implicit* equalization methods, which avoid the computation of a matrix inverse altogether.
- We propose a method that approximates the post-equalization signal-to-noise-and-interference-ratio (SINR) values in order to enable soft-output data detection with implicit equalizers.
- We propose two low-complexity initialization schemes that improve convergence of our iterative algorithms.
- We propose a hybrid explicit/implicit frequency-adaptive equalizer (short FADE) that exploits frequency correlation in wideband MIMO wireless systems and combines the advantages of explicit and implicit methods.

In order to evaluate the efficacy of our algorithms, we study the associated performance/complexity trade-offs in a 3GPP LTE-based massive MU-MIMO wireless system[1], and we provide a detailed performance and complexity comparison with existing approximate inversion algorithms proposed in [1, 6–11].

[1] The methods proposed in this paper can easily be extended to other multi-carrier waveforms that support frequency-domain equalization [13], such as OFDM-based systems.

1.2 Notation

Lowercase boldface letters stand for column vectors; uppercase boldface letters designate matrices. For a matrix $\mathbf{A}$, we denote its transpose and conjugate transpose $\mathbf{A}^T$ and $\mathbf{A}^H$, respectively. The entry in the k^{th} row and ℓ^{th} column of $\mathbf{A}$ is $A_{k,\ell}$; the k^{th} entry of a vector $\mathbf{a}$ is a_k. The Frobenius norm, the spectral norm, the ℓ_1-norm, and the ℓ_∞-norm of a matrix $\mathbf{A}$ are denoted by $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|_2$, $\|\mathbf{A}\|_1$, and $\|\mathbf{A}\|_\infty$, respectively. The $M \times M$ identity matrix is $\mathbf{I}_M$, $\mathbf{0}_{M\times N}$ is an $M \times N$ all-zeros matrix, and $\mathbf{F}_M$ refers to the orthonormal $M \times M$ discrete Fourier transform (DFT) matrix satisfying $\mathbf{F}_M^H \mathbf{F}_M = \mathbf{I}_M$.

1.3 Paper Outline

The rest of the paper is organized as follows. Section 2 details the system model and equalization-based data detection for SC-FDMA systems. Section 3 and Section 4 proposes explicit and implicit equalization methods, respectively. Section 5 and Section 6 proposes new initialization schemes and the frequency-adaptive equalizer (FADE), respectively. Section 7 and Section 8 provides a complexity comparison and a trade-off analysis, respectively. We conclude in Section 9.

2 System Model and Data Detection

We now introduce the LTE uplink model and detail an efficient, equalization-based minimum mean-square error (MMSE) data detector for SC-FDMA-based massive MU-MIMO systems. In what follows, we make frequent use of the superscript $(\cdot)^{(i,j)}$ to indicate the ith base-station antenna and the jth user; the subscript $(\cdot)_w$ designates the SC-FDMA subcarrier index.

2.1 Uplink System Model

We consider an LTE-based uplink system in which $U \leq B$ single-antenna[2] user terminals communicate with B BS antennas. The ith user first performs discrete Fourier transform (DFT) and subcarrier mapping for its serial time-domain (TD) data taken from a discrete constellation set $\mathcal{O}$ (e.g., QPSK) and generates the frequency domain (FD) symbol vector $\mathbf{s}^{(i)} = \left[s_1^{(i)}, \ldots, s_L^{(i)}\right]^T$. For each user, these FD symbols are assigned to data-carrying subcarriers, and then transformed back to the TD using an inverse discrete Fourier

[2] Our results can be extended to user terminals with multiple antennas.

transform (IDFT). After prepending the cyclic prefix, all U users transmit their TD signals over the wireless channel. The TD signals received at each BS antenna are first transformed back to the FD using a discrete Fourier transform (DFT), followed by the extraction of the data symbols. The received FD symbols on the wth subcarrier are modeled as $\mathbf{y}_w = \mathbf{H}_w\mathbf{s}_w + \mathbf{n}_w$ with

$$\mathbf{y}_w = \begin{bmatrix} y_w^{(1)} \\ \vdots \\ y_w^{(B)} \end{bmatrix}, \qquad \mathbf{H}_w = \begin{bmatrix} H_w^{(1,1)} & \cdots & H_w^{(1,U)} \\ \vdots & \ddots & \vdots \\ H_w^{(B,1)} & \cdots & H_w^{(B,U)} \end{bmatrix},$$

$$\mathbf{s}_w = [s_w^{(1)}, \ldots, s_w^{(U)}]^T, \quad \mathbf{n}_w = [n_w^{(1)}, \ldots, n_w^{(B)}]^T.$$

Here, $y_w^{(i)}$ is the FD symbol received on the wth subcarrier for antenna i, $H_w^{(i,j)}$ is the frequency gain (or attenuation) on the wth subcarrier between the ith receive antenna and jth user. The scalar $s_w^{(j)}$ denotes the symbol transmitted by the jth user on the w^{th} subcarrier; the scalar $n_w^{(i)}$ models i.i.d. circularly-symmetric complex Gaussian noise with variance N_0.

2.2 Soft-Output Data Detection

The goal of soft-output MIMO data detection is to generate reliability information in terms of LLR values for the transmitted data bits. Among the best-performing data-detection algorithms for traditional, small-scale MIMO systems are tree-search algorithms [14–16] (see [17] for a recent survey article). Unfortunately, such non-linear data-detection algorithms do not scale well to massive MU-MIMO systems with a large number of users and prevent efficient hardware designs. In recent years, alternative non-linear methods have been proposed for massive MU-MIMO systems, such as parallel interference cancellation [18,19], Monte–Carlo methods [20], and message-passing algorithms [21, 22]. All these methods exhibit significantly better complexity-scaling properties than tree-search methods, while enabling near-optimal performance with massive MU-MIMO. Nevertheless, none of these methods are directly applicable to SC-FDMA-based systems. For SC-FDMA, only a handful of non-linear data-detection methods exist [23–25]. While these methods achieve near-optimal performance in small-scale MIMO systems, their complexity does, similarly to tree-search-based methods, not scale well to large BS antenna arrays. In contrast, equalization-based linear data detection was shown in [6] to achieve near-optimal performance for SC-FDMA-based massive MU-MIMO systems, and the throughput of corresponding application-specific integrated circuits readily exceeds 3.8 Gb/s [12]. Thus, to achieve the throughputs required in future massive MU-MIMO systems at near-optimal performance, we focus on methods that rely on linear equalization.

2.3 Equalization-Based Linear Soft-Output Data Detection

The methods proposed in this paper build upon the MMSE data detector in [26] initially developed for traditional, small-scale MIMO-OFDM systems. This algorithm performs data-detection in two phases: (i) Estimates of the transmitted FD symbols in SC-FDMA systems are obtained using MMSE equalization on a per-subcarrier basis; (ii) log-likelihood ratio (LLR) values are computed in the time domain.

(i) Equalization: To perform MMSE equalization in the FD, we compute the Gram matrix $\mathbf{G}_w = \mathbf{H}_w^H\mathbf{H}_w$ and the matched filter vector $\mathbf{y}_w^{\text{MF}} = \mathbf{H}_w^H\mathbf{y}_w$ for each subcarrier w. We then compute the regularized Gram matrix $\mathbf{A}_w = \mathbf{G}_w + N_0E_s^{-1}\mathbf{I}_U$, which enables us to compute the equalized FD symbols as

$$\tilde{\mathbf{y}}_w = \mathbf{A}_w^{-1}\mathbf{y}_w^{\text{MF}}, \tag{1}$$

which are then used to compute the LLR values required for soft-output data detection [6, 26].

(ii) Soft-output Data Detection: Since the LTE uplink utilizes SC-FDMA, we first perform an IDFT on $\tilde{\mathbf{y}}^{(i)} = [\tilde{y}_1^{(i)}, \ldots, \tilde{y}_L^{(i)}]^T$ to obtain the TD estimate $\tilde{\mathbf{x}}^{(i)} = [\tilde{x}_1^{(i)}, \ldots, \tilde{x}_L^{(i)}]^T$. The so-called max-log LLR value of the j^{th} bit of t^{th} symbol, $L_{(t,j)}^{(i)}$, can then be computed as follows [26]:

$$L_{(t,j)}^{(i)} = \rho^{(i)}\left(\min_{a\in\mathcal{O}_j^0}\left|\frac{\tilde{x}_t^{(i)}}{\mu^{(i)}} - a\right|^2 - \min_{a\in\mathcal{O}_j^1}\left|\frac{\tilde{x}_t^{(i)}}{\mu^{(i)}} - a\right|^2\right). \tag{2}$$

Here, $\mathcal{O}_j^0$ and $\mathcal{O}_j^1$ are the constellation subsets for which the j bit is 0 and 1 respectively. The post-equalization signal-to-noise-and-interfence-ratio (SINR) is $\rho^{(i)} = (\mu^{(i)})^2/\nu_i^2$, where $\nu_i^2 = E_s\mu^{(i)} - E_s|\mu^{(i)}|^2$ for SC-FDMA-based systems. The effective channel gain is computed as $\mu^{(i)} = L^{-1}\sum_{w=1}^{L}\mathbf{a}_{i,w}^H\mathbf{g}_{i,w}$, where $\mathbf{a}_{i,w}^H$ is the i^{th} row of $\mathbf{A}_w^{-1}$ and $\mathbf{g}_{i,w}$ is the i^{th} column of $\mathbf{G}_w$. See [6] for more details.

2.4 Explicit vs. Implicit Equalization

There exist two distinct equalization methods to compute (1), namely *explicit* and *implicit* methods. Explicit methods first compute the matrix inverse $\mathbf{A}_w^{-1}$ (or an approximation thereof) and then, use the matrix inverse to compute the equalized FD symbol as in (1). Implicit methods solve the system of linear equations

$\mathbf{A}_w\tilde{\mathbf{y}}_w = \mathbf{y}_w^{\mathrm{MF}}$ either exactly or approximately to compute the equalized FD symbol $\tilde{\mathbf{y}}_w$; this approach avoids an explicit computation of the inverse $\mathbf{A}_w^{-1}$.

The key advantage of implicit equalization methods is the fact that they require (often significantly) lower computational complexity than explicit methods. In contrast, explicit equalization methods have the following advantages: (i) Massive MU-MIMO systems are expected to operate as time-division duplexing (TDD) systems [1], in which the BS estimates the channel during the uplink phase. As a result, the matrix inverse obtained during the uplink transmission can be re-used to perform MU precoding (or beamforming) in the downlink. (ii) For slow-fading channels and/or flat-fading channels with low-delay spread, the inverse can be re-used for consecutive symbols and/or adjacent subcarriers, respectively. (iii) Computation of the post-equalization SINR $\rho^{(i)}$ used in the LLR computation (2) can be obtained from the explicit inverse $\mathbf{A}_w^{-1}$ (see Section 2.3). In the following Sections 3 and 4, we discuss both of these equalization schemes.

3 Explicit Equalization

We start by discussing *explicit* MMSE equalization, i.e., where we obtain the equalized symbol $\tilde{\mathbf{y}}_w$ by first computing or approximating the inverse matrix $\mathbf{A}_w^{-1}$, followed by computing (1). We provide an overview of exact, explicit inversion methods and proceed by discussing existing and new iterative methods that approximate $\mathbf{A}_w^{-1}$ at low complexity. To simplify notation, we omit the subcarrier index w.

3.1 Exact Inversion via the Cholesky Decomposition

The literature describes a large number of exact methods to compute $\mathbf{A}^{-1}$; see the references [27–29] for an overview. One of the most efficient methods (in terms of arithmetic operations) that can be implemented in VLSI at low complexity relies on the Cholesky decomposition [6, 30–32]. This approach first factorizes the regularized Gram matrix $\mathbf{A} = \mathbf{L}\mathbf{L}^H$, where $\mathbf{L}$ is a lower-triangular matrix with non-negative entries on the main diagonal. To obtain $\mathbf{A}^{-1}$, this approach then solves $\mathbf{L}\mathbf{X} = \mathbf{I}_U$ for $\mathbf{X}$ using forward substitution—one can then solve $\mathbf{L}^H\mathbf{A}^{-1} = \mathbf{X}$ for $\mathbf{A}^{-1}$ using back substitution.

The complexity that is required to explicitly compute $\mathbf{A}^{-1}$ using the Cholesky decomposition can become prohibitive for large values of U (see Section 7). Furthermore, computing the Cholesky decomposition, as well as performing forward or backward substitution, exhibits stringent data dependencies, which prevents highly-parallel hardware architectures. To reduce the computational complexity of matrix inversion for the high-dimensional systems anticipated in systems that use massive MU-MIMO and to enable massively parallel hardware designs, we next propose novel, low-complexity methods that explicitly compute approximate versions of the matrix inverse $\mathbf{A}^{-1}$.

3.2 Exact Inversion using Series Expansions

3.2.1 Accelerated Neumann Series Expansion

In [6], the authors proposed a truncated version of the Neumann series expansion [33] with the goal of reducing the complexity of exact, explicit matrix inversion. We now propose a more general, accelerated version of the classical Neumann series, which enables the design of approximate data detectors that achieve superior error-rate performance at low complexity. The proof of the following result is given in Appendix A.1.

Lemma 1 (Accelerated Neumann Series)**.** *Let* $\widetilde{\mathbf{A}}_0^{-1} \in \mathbb{C}^{U\times U}$ *be a so-called* initialization matrix *with full rank. Suppose that*

$$\lim_{k\to\infty} (\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A})^k = \mathbf{0}_{U\times U}. \tag{3}$$

Then, we have the following accelerated Neumann series*:*

$$\mathbf{A}^{-1} = \textstyle\sum_{k=0}^{\infty}(\mathbf{I} - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A})^k\widetilde{\mathbf{A}}_0^{-1}. \tag{4}$$

In Section 5, we will develop efficient methods for computing initialization matrices $\widetilde{\mathbf{A}}_0^{-1}$, which enable accurate approximations of $\mathbf{A}^{-1}$ with only a few terms of the accelerated Neumann series in (4). In order to design such matrices, we will make use of the following convergence condition; the proof directly follows from [33, Thm. 4.20].

Lemma 2 (Convergence Condition)**.** *A sufficient condition for (3) to hold is that*

$$\|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\| < 1 \tag{5}$$

for any consistent matrix norm.

3.2.2 Accelerated Neumann Series Recursion

As it will be important for the implicit, approximate inversion methods discussed in Section 4, it is key to realize that (4) can alternatively be formulated using the following recursion for the iterations $k = 1, 2, \ldots$ given by

$$\widetilde{\mathbf{A}}_k^{-1} = \widetilde{\mathbf{A}}_0^{-1} + \big(\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\big)\widetilde{\mathbf{A}}_{k-1}^{-1}, \tag{6}$$

which we initialize with $\widetilde{\mathbf{A}}_0^{-1}$ (hence, the name initialization matrix). Given that (5) holds, the recursion satisfies $\lim_{k\to\infty} \widetilde{\mathbf{A}}_k^{-1} = \mathbf{A}^{-1}$. The recursion in (6) can be derived from the right-hand side (RHS) in (4) by successively factoring $\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}$ from the infinite sum. We note that recurrent operations in (6) can be avoided in practice by precomputing the matrix $\widetilde{\mathbf{A}}_0^{-1}\mathbf{A}$.

3.2.3 Schulz Recursion

To obtain faster convergence rates than the accelerated Neumann series recursion (6), one may use higher-order recursions. One prominent method is the Schulz recursion [34], which has been proposed for small-scale MIMO systems in [35]. As for (6), if (5) holds, then the inverse $\mathbf{A}^{-1}$ can be computed recursively for $k = 1, 2, \ldots$ as follows [34]:

$$\widetilde{\mathbf{A}}_k^{-1} = \big(2\,\mathbf{I}_U - \widetilde{\mathbf{A}}_{k-1}^{-1}\mathbf{A}\big)\widetilde{\mathbf{A}}_{k-1}^{-1} \tag{7}$$

with the initialization matrix $\widetilde{\mathbf{A}}_0^{-1}$. This recursion generates 2^k Neumann series terms for k iterations, whereas the accelerated Neumann recursion (6) only generates $k+1$ terms per k iterations. The Schulz method (7), however, requires two matrix multiplications per iteration, in contrast to the accelerated Neumann recursion (6) that requires only one. Hence, for a small number of iterations, i.e., for $k \leq 2$, the accelerated Neumann series recursion is computationally more efficient than the Schulz recursion (see Section 7.2 for more details).

3.2.4 Higher-Order Recursions

The literature describes other recursive inversion methods [36, 37], which converge even faster than the Schulz recursion (7). For example, if (5) holds, then the inverse $\mathbf{A}^{-1}$ can be computed recursively for $k = 1, 2, \ldots$ as follows [37]:

$$\widetilde{\mathbf{A}}_k^{-1} = \widetilde{\mathbf{A}}_{k-1}^{-1}\Big(3\,\mathbf{I}_U - \mathbf{A}\widetilde{\mathbf{A}}_{k-1}^{-1}\big(3\,\mathbf{I}_U - \mathbf{A}\widetilde{\mathbf{A}}_{k-1}^{-1}\big)\Big) \tag{8}$$

with the initialization matrix $\widetilde{\mathbf{A}}_0^{-1}$. This 3rd-order matrix inverse approximation requires three matrix multiplications per iteration (if one precomputes $\mathbf{A}\widetilde{\mathbf{A}}_{n-1}^{-1}$) and generates 3^k Neumann series terms for k iterations. Note that (8) is computationally more efficient than the Schulz recursion only for $k \geq 3$.

3.3 Approximate Inversion using Truncated Series Expansions

For a large number of iterations, computing the accelerated Neumann Series recursion (6), as well as the recursions in (7) and (8), is impractical and entails higher complexity than the Cholesky-based approach in Section 3.1. However, if we restrict ourselves to a small number $K_{\max}$ of iterations, which is equivalent to *truncating* the series (4) to $K_{\max}$ terms, one can accurately approximate $\mathbf{A}^{-1}$ at low complexity. We next discuss existing and new variations of this general idea.

3.3.1 Truncated Neumann Series

The approximate, explicit inversion approach put forward in [6] evaluates only $K_{\max}$ terms in (4) together with a simple initialization matrix. Reference [6] starts by decomposing the matrix $\mathbf{A}$ into its main diagonal part $\mathbf{D}$ and the off-diagonal part $\mathbf{E} = \mathbf{A} - \mathbf{D}$. Then, by using $\widetilde{\mathbf{A}}_0^{-1} = \mathbf{D}^{-1}$ as the initialization matrix, the following truncated Neumann series [6]

$$\widetilde{\mathbf{A}}_{K_{\max}}^{-1} = \textstyle\sum_{k=0}^{K_{\max}} (-\mathbf{D}^{-1}\mathbf{E})^k \mathbf{D}^{-1} \tag{9}$$

accurately approximates $\mathbf{A}^{-1}$ for small values of $K_{\max}$ in systems with large BS-to-user-antenna ratios.

For $K_{\max} = 0$, this approach results in $\widetilde{\mathbf{A}}_0^{-1} = \mathbf{D}^{-1}$, which, together with (1), results in a scaled version of the matched filter (MF) equalizer $\tilde{\mathbf{y}} = \mathbf{D}^{-1}\mathbf{y}^{\mathrm{MF}}$. For slightly larger values of $K_{\max}$ (e.g., one or two), we can trade-off performance versus complexity. In fact, the complexity of this approximation is quadratic and cubic for $K_{\max} = 1$ and $K_{\max} = 2$ respectively, while $K_{\max} = 2$ outperforms $K_{\max} = 1$ in terms of the error rate (see [6] for a detailed trade-off analysis). We note that the convergence condition (5) is not guaranteed[3] to hold for the choice $\widetilde{\mathbf{A}}_0^{-1} = \mathbf{D}^{-1}$. In Section 5, we will develop novel choices for the initialization matrix $\widetilde{\mathbf{A}}_0^{-1}$ that not only require low complexity but also yield more accurate approximates of $\mathbf{A}^{-1}$.

3.3.2 Higher-Order Series Expansions

Evidently, the above truncation approach can also be used in combination with the Schulz recursion (7) or other higher-order recursions, such as the one in (8). In Sections 7 and 8 we will analyze the associated performance/complexity trade-offs.

3.4 LLR Computation for Explicit Inversion Methods

With the above methods for computing the inverse $\mathbf{A}_w^{-1}$, we can calculate the LLR values for soft-output data detection using (2). Specifically, for the Cholesky decomposition in Section 3.1 and the exact series expansions in Section 3.2, we first compute the equalized symbols (1)

[3] A probabilistic convergence condition is provided in [6].

and then, generate the TD estimates $\tilde{\mathbf{x}}_w$ with an IDFT. The LLR values (2) are obtained directly from $\mathbf{A}_w^{-1}$ and $\tilde{\mathbf{x}}_w$, where the quantities $\mu_w^{(i)}$ and $\rho_w^{(i)}$ are computed as discussed in Section 2.3.

For the approximate matrix inversion methods described in Section 3.3, we first compute the approximate equalized symbol $\tilde{\mathbf{y}}_w = \mathbf{A}_{w|K_{\max}}^{-1}\mathbf{y}_w^{\mathrm{MF}}$ for each subcarrier and then, generate the TD estimates $\tilde{\mathbf{x}}_w$ with an IDFT. To obtain LLR values (2), we first compute $\mu_K^{(i)} = L^{-1}\sum_{w=1}^{L}\tilde{\mathbf{a}}_{i,w|K}^H\mathbf{g}_{i,w}$, where $\tilde{\mathbf{a}}_{i,w|K}^H$ is ith row of $\mathbf{A}_{w|K}^{-1}$. The missing quantity, $\rho^{(i)} = (\mu^{(i)})^2/\nu_i^2$, however, requires a matrix-matrix multiplication as we need to compute [6]:

$$\nu^{(i)} = E_s\sum_{w=1}^{L}\tilde{\mathbf{a}}_{i,w|K}^H\mathbf{A}_w\mathbf{G}_w\tilde{\mathbf{a}}_{i,w|K} - E_s|\mu_K^{(i)}|^2. \quad (10)$$

Computing this expression requires the same order of complexity (i.e., cubic) as the approximate matrix inverse itself and hence, should be avoided to maintain low complexity. To this end, one can use the approximation proposed in [6], which uses $\nu_i^2 \approx E_s\mu_0^{(i)} - E_s|\mu_0^{(i)}|^2$ with $\mu_0^{(i)} = L^{-1}\sum_{w=1}^{L}(d_w^{(i)})^{-1}g_w^{(i)}$. Note that this approximation approaches its exact counterpart in the large-antenna limit for massive MU-MIMO systems [6].

4 Implicit Equalization

We now discuss existing and novel *implicit* MMSE equalization algorithms. The idea of these methods is to obtain the equalized symbol $\tilde{\mathbf{y}}_w$ (or a corresponding approximation) without ever computing the inverse $\mathbf{A}_w^{-1}$. As discussed in Section 2.4, the complexity of implicit methods is, in general, significantly lower than for explicit methods. Nevertheless, *exact* computation of the post-equalization SINR, as required for LLR computation (2), is computationally expensive. To enable soft-output data detection with implicit equalization methods, we propose a low-complexity SINR approximation. To simplify notation, we omit the subcarrier index w.

4.1 Exact Inversion using Implicit Cholesky Decomposition

Implicit equalization methods solve for $\tilde{\mathbf{y}}$ directly without computing $\mathbf{A}^{-1}$ explicitly. A hardware-friendly approach for implicit equalization first performs the Cholesky decomposition to obtain $\mathbf{A} = \mathbf{L}\mathbf{L}^H$. Then, one can solve $\mathbf{L}\mathbf{x} = \mathbf{y}^{\mathrm{MF}}$ for $\mathbf{x}$ followed by solving $\mathbf{L}^H\tilde{\mathbf{y}} = \mathbf{x}$, where $\tilde{\mathbf{y}}$ corresponds to the equalized vector (see, e.g., [32] for a corresponding hardware design).

4.2 Implicit Accelerated Neumann Recursion

To reduce the complexity of implicit equalization, we can perform the following implicit, accelerated Neumann recursion; the proof immediately follows from right-multiplying both sides of (6) by $\mathbf{y}^{\mathrm{MF}}$.

Lemma 3 (Implicit Accelerated Neumann Recursion)**.** *Let $\tilde{\mathbf{y}}_0 = \widetilde{\mathbf{A}}_0^{-1}\mathbf{y}^{MF}$, where the initialization matrix $\widetilde{\mathbf{A}}_0^{-1}$ satisfies (5). Then, for the iterations $k = 1, 2, \ldots$*

$$\tilde{\mathbf{y}}_k = \tilde{\mathbf{y}}_0 + \big(\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\big)\tilde{\mathbf{y}}_{k-1} \quad (11)$$

we recursively obtain $\tilde{\mathbf{y}}_k = \widetilde{\mathbf{A}}_k^{-1}\mathbf{y}^{MF}$ with $\widetilde{\mathbf{A}}_k^{-1}$ as defined in (6).

Evidently, the recursion (11) can be terminated after $K_{\max}$ iterations to obtain an approximate to (1) at low complexity. Furthermore, $\widetilde{\mathbf{A}}^{-1}\mathbf{A}$ can be precomputed in practice to avoid recurrent calculations. We note that the recursion in (11) is a generalization of the equalization algorithm proposed in [1], which uses $\widetilde{\mathbf{A}}_0^{-1} = \mathbf{I}_U$. Note that this particular choice only performs well for suitably normalized channel matrices and massive MU-MIMO systems with large BS-to-user-antenna ratios.

Unfortunately, the Schulz recursion and higher-order recursions do not—to the best of our knowledge—have efficient implicit forms. In fact, if we right-multiply both sides of (7) or (8) by $\mathbf{y}^{\mathrm{MF}}$, we see that one needs to keep track of the matrix $\widetilde{\mathbf{A}}_k^{-1}$ in order to compute $\tilde{\mathbf{y}}_k$; this prevents the design of a computationally efficient, implicit recursion with these methods.

4.3 Existing Approximate Implicit Equalization Methods

A variety of low-complexity, implicit equalization methods for data detection in massive MU-MIMO systems have been proposed recently [8–10, 38]. The Richardson method proposed in [9] can be rewritten as

$$\tilde{\mathbf{y}}_k = \bar{\gamma}\mathbf{y}^{\mathrm{MF}} + (\mathbf{I}_U - \bar{\gamma}\mathbf{A})\,\tilde{\mathbf{y}}_{k-1},$$

which corresponds to a special case of the accelerated implicit Neumann series recursion in (11) with $\widetilde{\mathbf{A}}_0^{-1} = \bar{\gamma}\mathbf{I}$; the quantity $\bar{\gamma}$ is an algorithm parameter.[4] Other implicit methods, such as the conjugate gradient method (CG) method [10] and the Gauss-Seidel (GS) algorithm [8] are iterative methods that solve systems of linear equations for the positive semidefinite matrix $\mathbf{A}$. Both methods, CG and GS, will converge to the exact solution for a sufficiently large number of iterations. GS is initialized by $\tilde{\mathbf{y}}_0 = \mathbf{D}^{-1}\mathbf{y}^{\mathrm{MF}}$; for CG, we define the

[4] We note that the parameter in [9] is $\bar{\gamma} = (B+U)^{-1}$.

vector $\tilde{\mathbf{y}}_0$ as the output of the first iteration since the initial guess is an all-zero vector. CG and GS enable approximate equalization at (often) lower complexity than other explicit and implicit equalization algorithms [8–10]. In Sections 7 and 8 we compare the computational complexity and performance of these equalization methods, respectively.

4.4 LLR Approximation for Implicit Inversion Methods

We can evaluate (2) to obtain LLR values for the transmitted bits. Since the proposed implicit methods do not compute the matrix inverse $\mathbf{A}_w^{-1}$ (or a corresponding approximation $\mathbf{A}_{w|K_{\max}}$), computing the quantities $\mu^{(i)}$ and $\rho^{(i)}$ without having $\mathbf{A}_w^{-1}$ seems difficult. To enable soft-output data detection with implicit equalization algorithms, we propose a novel approximation for $\mu^{(i)}$ and $\rho^{(i)}$ that does not need the explicit inverse $\mathbf{A}_w^{-1}$ or a corresponding approximation $\mathbf{A}_{w|K_{\max}}^{-1}$.

Our approach uses the effective channel gain $\mu^{(i)}$ for the 0^{th}-term Neumann series approximation

$$\mu^{(i)} \approx \tilde{\mu}_0^{(i)} = L^{-1} \textstyle\sum_{w=1}^{L} (d_w^{(i)})^{-1} g_w^{(i)}, \tag{12}$$

where $d_w^{(i)}$ is the i^{th} diagonal element of $\mathbf{A}_w$ and $g_w^{(i)}$ is the i^{th} diagonal element of $\mathbf{G}_w$. Analogously, we propose to use the SINR for the 0^{th}-term Neumann series approximation

$$\mu^{(i)} \approx \tilde{\rho}_0^{(i)} = (\mu_0^{(i)})^2/(\nu_0^{(i)})^2 = \tilde{\mu}_0^{(i)} \big(E_s \tilde{\mu}_0^{(i)} - E_s |\tilde{\mu}_0^{(i)}|^2\big)^{-1}.$$

As we will demonstrate in Section 8, this LLR approximation enables implicit equalizers that achieve near-optimal error-rate performance at low complexity.

5 Initialization Matrices

The proposed series-based explicit and implicit methods in Sections 3 and 4 require a suitable initialization matrix $\widetilde{\mathbf{A}}_0^{-1}$ to improve (i) the probability of convergence, i.e., the probability that the initialization matrix satisfies (5), and (ii) the accuracy of the approximated matrix inverse when performing only a small number of iterations. We next discuss existing choices for $\widetilde{\mathbf{A}}_0^{-1}$ and propose two new methods that enable improved error-rate performance.

5.1 Relevance of the Initialization Matrix

We first show that the choice of the initialization matrix $\widetilde{\mathbf{A}}^{-1}$ directly affects the performance of (explicit and implicit) approximate equalizers that use truncated series expansions. The following Lemma generalizes the derivations in [6, Sec. III-B] for the initializer $\mathbf{D}^{-1}$; a short proof is given in Appendix A.2.

Lemma 4 (Residual Estimation Error)**.** *Let* $\tilde{\mathbf{y}}_{K_{max}} = \mathbf{A}_{K_{max}}^{-1}\mathbf{y}^{MF}$ *be the result of an approximate equalizer using a truncated series expansion. Define the residual estimation error as*

$$\mathbf{e}_{K_{max}} = \tilde{\mathbf{y}}_{K_{max}} - \mathbf{A}^{-1}\mathbf{y}^{MF}.$$

Then, we have the following upper bound on the residual estimation error:

$$\|\mathbf{e}_{K_{max}}\| \leq \|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|^{K_{max}+1}\|\tilde{\mathbf{y}}\|,$$

where $\tilde{\mathbf{y}} = \mathbf{A}^{-1}\mathbf{y}^{MF}$ *is the estimate obtained through the exact equalizer and* $\|\cdot\|$ *is a consistent (matrix) norm.*

It is evident that by reducing $\|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|$, we directly reduce the residual estimation error. Furthermore, if (5) holds, then increasing the number of accelerated Neumann series terms $k \to \infty$ forces the residual estimation error to zero, i.e., the series expansion is exact. Hence, it is of utmost important to chose an initialization matrix $\widetilde{\mathbf{A}}_0^{-1}$ that minimizes $\|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|$ in order to minimize both the residual error and, consequently, the error-rate of approximate linear equalization.

5.2 Existing Initialization Matrices

Common initialization matrices [36, 37, 39]. that satisfy (5) are of the form $\widetilde{\mathbf{A}}_0^{-1} = \alpha\mathbf{A}^H$, where $\alpha > 0$ is a carefully chosen scalar. For example, reference [37] postulates the use of $\alpha^{-1} = (\lambda_{\max} + \lambda_{\min})/2$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of $\mathbf{A}^H\mathbf{A}$, respectively. This choice minimizes the left-hand side of (5) by assuming the spectral norm. Unfortunately, the complexity required to compute the largest and smallest eigenvalues of $\mathbf{A}^H\mathbf{A}$ is, in our application, larger than computing the inverse $\mathbf{A}^{-1}$ itself, which renders this method unattractive. Related approaches that ensure (5) while requiring lower complexity are, for example, $\alpha^{-1} = \frac{1}{2}\|\mathbf{A}\mathbf{A}^H\|_\infty$ or $\alpha^{-1} = \|\mathbf{A}\|_1\|\mathbf{A}\|_\infty$ [36, 39].

Reference [6] proposes $\widetilde{\mathbf{A}}_0^{-1} = \mathbf{D}^{-1}$, where $\mathbf{D}$ is the main diagonal of $\mathbf{A}$. While this initialization approach requires low complexity and was shown to perform well for data detection in massive MU-MIMO systems, it does not guarantee (5) to hold. Nevertheless, as shown in [6, Thm. 1], for i.i.d. circularly-symmetric complex Gaussian channel matrices $\mathbf{H}$ and for sufficiently large BS-to-user-antenna ratios (e.g., two or higher), the condition (5) is satisfied with high probability. Unfortunately, for small BS-to-user-antenna ratios, this initialization method results in poor error-rate performance.

The more recent reference [9] proposes $\widetilde{\mathbf{A}}_0^{-1} = (U + B)^{-1}\mathbf{I}_U$, which only converges in the large antenna limit for i.i.d. circularly symmetric complex Gaussian channel matrices $\mathbf{H}$. This method, however, still diverges for "not-so-massive systems" with small BS-to-user-antenna ratios (see Section 5.4).

5.3 Two New Initialization Matrices

We next propose two new initialization matrices that can be computed at low complexity and result in small approximation errors even for a few iterations $K_{\max}$. In addition, as we will show in Section 5.4, the proposed initializers outperform the methods discussed in Section 5.2 in practical scenarios. We emphasize that the initialization matrices proposed next are suitable for any matrix inversion method that uses (truncated) series expansions, i.e., for applications beyond equalization and data detection.

The first initialization method requires low complexity; see Section 7.1 for a discussion. In Appendix A.3, we derive this initialization method by minimizing $\alpha \in \mathbb{C}$ in condition (5) for matrices of the form $\widetilde{\mathbf{A}}_0^{-1} = \alpha\mathbf{D}^{-1}$.

Initialization 1. *Let* $\mathbf{D}$ *contain the main diagonal of* $\mathbf{A}$*. Then, the initialization* $\widetilde{\mathbf{A}}_0^{-1} = \alpha^{opt}\mathbf{D}^{-1}$ *with* $\alpha^{opt} = U\|\mathbf{D}^{-1}\mathbf{A}\|_F^{-2}$ *minimizes (5).*

The second initialization scheme refines Initialization 1 at slightly higher computational complexity; see Section 7.1 for a discussion. The method is derived in Appendix A.4, where we minimize the parameters $\{\alpha, \ldots, \alpha_U\}$ in condition (5) for matrices of the form $\widetilde{\mathbf{A}}_0^{-1} = \mathrm{diag}(\alpha_1, \ldots, \alpha_U)\mathbf{D}^{-1}$.

Initialization 2. *Let* $\mathbf{D}$ *be the main diagonal of* $\mathbf{A}$*. Furthermore, let* $\mathbf{Q}$ *contain the off-diagonal part of* $\mathbf{D}^{-1}\mathbf{A}$*. Then, the initialization* $\widetilde{\mathbf{A}}_0^{-1} = diag(\alpha_1^{opt}, \ldots, \alpha_U^{opt})\mathbf{D}^{-1}$ *with* $\alpha_i^{opt} = (1 + \|\mathbf{r}_i\|_2^2)^{-1}$*, where* $\mathbf{r}_i$ *is the ith row of* $\mathbf{Q}$*, minimizes (5).*

We conclude by noting that both of these initialization schemes do not, in general, guarantee convergence according to (5) as we only minimized the free parameters α or $\{\alpha_1, \ldots, \alpha_U\}$. Nevertheless, the proposed initialization methods exhibit (often significantly) faster convergence compared to the ones discussed in Section 5.2 and converge (empirically) with high probability[5], even for small BS-to-user-antenna ratios. We next empirically study the convergence behavior of all the discussed and proposed initialization matrices.

[5] The derivation of probabilistic convergence guarantees turns out to be non-trivial and is part of ongoing work.

5.4 Comparison of Empirical Convergence Behavior

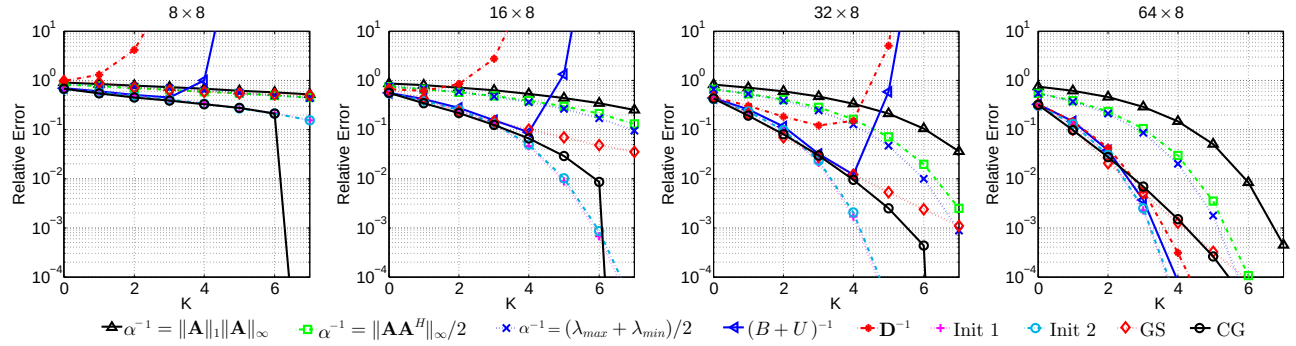
To assess the convergence behavior of approximate equalizers using the truncated series expansions for different initialization matrices, we generate $B \times U$ random matrices $\mathbf{H}$ and U-dimensional vectors $\mathbf{x}$, where the entries are i.i.d. circularly-symmetric complex Gaussian with unit variance. For each matrix and vector pair, we compute $\mathbf{y} = \mathbf{H}\mathbf{x}$. Given $\mathbf{H}$ and $\mathbf{y}$, we first compute the Gram matrix $\mathbf{A} = \mathbf{H}^H\mathbf{H}$ and perform recursive matrix inversion using the accelerated Neumann recursion (6), the Schulz recursion (7), and the 3rd order recursion (8). We then use the approximate inverse $\widetilde{\mathbf{A}}_k^{-1}$ to obtain $\widetilde{\mathbf{x}}_k = \widetilde{\mathbf{A}}_k^{-1}\mathbf{y}$. In addition, we also estimate $\widetilde{\mathbf{x}}_k$ using CG [10] and GS [8], which are both implicit methods.

Figure 1 compares the relative error (RE) at iteration k defined as $RE(k) = \|\mathbf{x} - \widetilde{\mathbf{x}}_k\|_2/\|\mathbf{x}\|_2$ between the exact and the approximate solution for the Neumann (Figure 1(a)), Schulz (Figure 1(b)), and 3rd order recursion (Figure 1(c)). We report the average RE over $10,000$ Monte-Carlo trials.

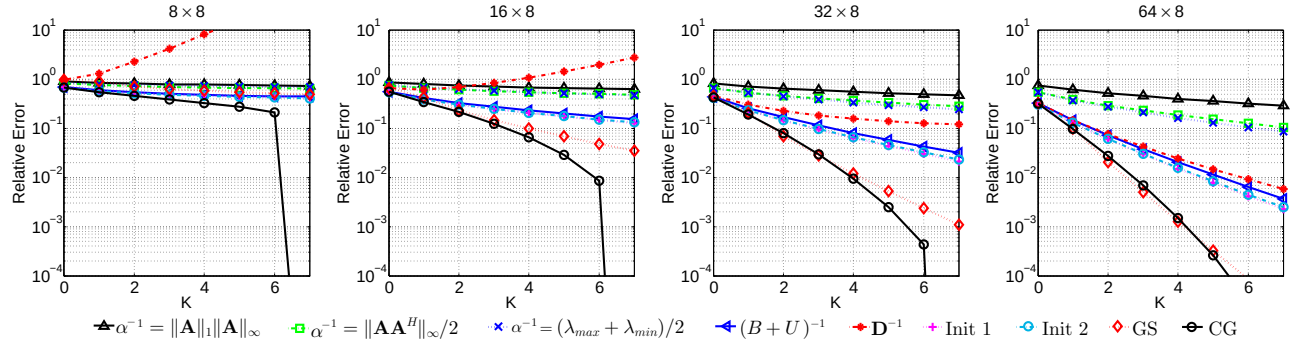
By comparing the convergence behavior of the Neumann recursion (Figure 1(a)) against the Schulz recursion (Figure 1(b)), and the 3rd order recursion (Figure 1(c)), we see that the average RE decreases faster for higher-order recursions. Although CG and GS outperform the methods that use a truncated Neumann series, they are both implicit methods and do not compute an approximate matrix inverse.

Figure 1 also shows the performance of conventional initialization matrices. Traditional initializers, such as $\widetilde{\mathbf{A}}_0^{-1} = \alpha\mathbf{A}^H$ with (i) $\alpha = 2(\lambda_{\max} + \lambda_{\min})^{-1}$, (ii) $\alpha = 2\|\mathbf{A}^H\|_\infty^{-1}$, and (iii) $\alpha = (\|\mathbf{A}\|_1\|\mathbf{A}\|_\infty)^{-1}$ always converge, whereas (i) leads to the fastest convergence among these methods. The initialization matrix $\mathbf{D}^{-1}$ proposed specifically for massive MU-MIMO leads to faster convergence than these traditional initialization matrices for large BS-to-user-antenna ratios but diverges as $k \to \infty$ for small ratios. The Richardson method [9] uses $\widetilde{\mathbf{A}}_0^{-1} = (B + U)^{-1}\mathbf{I}_U$ and exhibits similar convergence as $\mathbf{D}^{-1}$ and also diverges for small BS-to-user-antenna ratios. The proposed Initializers 1 and 2 enable faster convergence than all the other considered initialization matrices for all BS-to-user-antenna ratios.

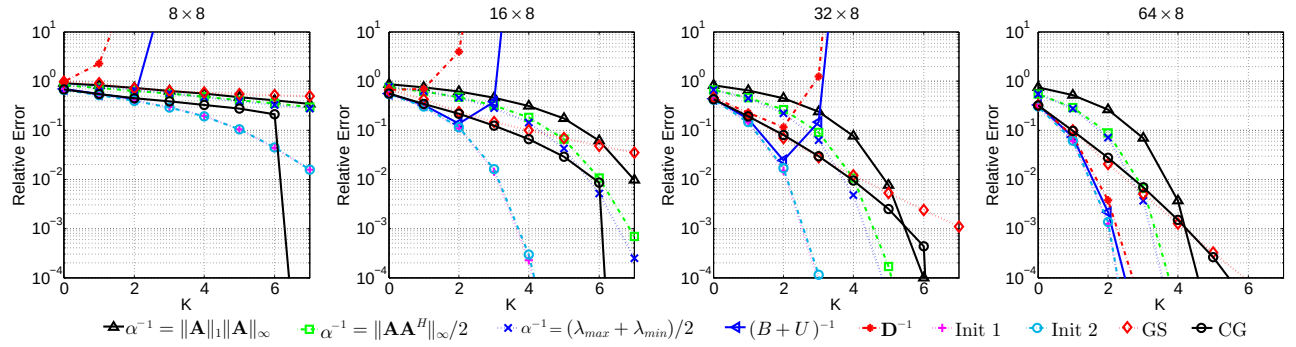
As a general trend, we observe that improved convergence is obtained for all considered algorithms in Figure 1 by increasing the BS-to-user-antenna ratio. This fact shows that massive MIMO enables low-complexity data detection methods to achieve near-optimal error-rate performance.



(a) Average RE of the truncated Neumann series recursion with different initializers.



(b) Average RE of the truncated Schulz recursion with different initializers.



(c) Average RE of the truncated 3rd order recursion with different initializers.

Fig. 1 Average relative error (RE) comparison for different antenna configuration, algorithms, and initialization methods. The proposed Initializers 1 and 2 outperform existing initializers, even for small BS-to-user-antenna ratios; CG and GS exhibit excellent convergence, whereas CG is exact for eight iterations (as we consider an eight-user system).

6 Frequency-Adaptive Equalizer (FADE)

We now propose the frequency adaptive equalizer (FADE, for short), which combines the advantages of implicit, explicit, exact, and approximate equalization methods, and achieves near-exact equalization performance at very low computational complexity.

6.1 Exploiting Correlation in Multipath Wireless Channels

Practical multipath channels in wideband communication systems typically exhibit correlation across time and frequency [40]. In fact, by assuming a wide-sense stationary uncorrelated scattering (WSSUS) channel model, the TD correlation between symbols is dependent on the Doppler spread and the FD correlation between subcarriers is dependent on the delay spread [41]. Existing wideband systems that rely on OFDM and SC-FDMA already exploit time and frequency correlation to estimate the channel coefficients. For example, 3GPP LTE-A [42] embeds pilot symbols across frequency and time in the transmitted signal, which allows the receiver to estimate the channel coefficients by means of interpolation.

FD correlation has also been exploited to reduce the complexity for linear data detection in traditional small-

scale MIMO systems [43, 44]. Reference [43] proposes to compute an explicit inverse only at a given number of subcarriers (so-called base-points), whereas the other inverses at the remaining subcarriers are computed through interpolation. Given a sufficiently large number of base points (depending on the delay spread), this method was shown to be exact. The drawback of such exact, interpolation-based matrix inversion methods for massive MU-MIMO is the high computational complexity caused by rather long interpolation filters [44]. Nevertheless, inspired by these algorithms, we next propose an approximate interpolation-based equalization method that achieves excellent performance at low complexity.

6.2 Frequency Adaptive Equalizer (FADE)

The key idea of FADE is to exploit correlation across frequency (and possibly time) and to take advantage of explicit and implicit equalization schemes. As in [43, 44], we first compute an explicit matrix inverse at a given (small) set Ω of subcarriers (base-points).[6] Given the inverse matrix $\mathbf{A}_w^{-1}$ at base-point $w \in \Omega$, we can approximate the matrix inverse at nearby (adjacent) subcarriers w' by using one of the accelerated recursions in Section 3.2 with $\mathbf{A}_w^{-1}$ as the initialization matrix (i.e. $\widetilde{\mathbf{A}}_{w'|0}^{-1} = \mathbf{A}_w^{-1}$). For example, the matrix inverse $\mathbf{A}_{w'}^{-1}$ at a neighboring subcarrier $w' = w+1$ can be approximated by computing one explicit recursion of the Neumann series in (11) using $\widetilde{\mathbf{A}}_{w+1|0}^{-1} = \mathbf{A}_w^{-1}$ as follows:

$$\begin{aligned}\widetilde{\mathbf{A}}_{w+1|1}^{-1} &= \mathbf{A}_w^{-1} + \left(\mathbf{I}_U - \mathbf{A}_w^{-1}\mathbf{A}_{w+1}\right)\mathbf{A}_w^{-1} \\ &= 2\mathbf{A}_w^{-1} - \mathbf{A}_w^{-1}\mathbf{A}_{w+1}\mathbf{A}_w^{-1}. \end{aligned} \quad (13)$$

Unfortunately, the complexity of (13) is dominated by two matrix multiplications, which is higher than that of a Cholesky decomposition. To reduce the complexity, we perform implicit equalization on neighboring subcarriers instead, i.e., we compute

$$\tilde{\mathbf{y}}_{w+1} = 2\mathbf{A}_w^{-1}\mathbf{y}_{w+1}^{\mathrm{MF}} - \mathbf{A}_w^{-1}\mathbf{A}_{w+1}\mathbf{A}_w^{-1}\mathbf{y}_{w+1}^{\mathrm{MF}}, \quad (14)$$

where $\mathbf{y}_{w+1}^{\mathrm{MF}} = \mathbf{H}_{w+1}^H\mathbf{y}_{w+1}$. Besides computationally-efficient matrix vector products, this implicit, approximate equalizer still requires computation of the regularized Gram matrix $\mathbf{A}_{w+1}$. It is, however, crucial to realize that the Gram matrix does not need to be computed when expanding (14) into

$$\begin{aligned}\tilde{\mathbf{y}}_{w+1} = {} & 2\mathbf{A}_w^{-1}\mathbf{y}_{w+1}^{\mathrm{MF}} - \\ & \mathbf{A}_w^{-1}\left(\mathbf{H}_{w+1}^H\mathbf{H}_{w+1}\mathbf{A}_w^{-1}\mathbf{y}_{w+1}^{\mathrm{MF}} + N_0E_s^{-1}\mathbf{A}_w^{-1}\mathbf{y}_{w+1}^{\mathrm{MF}}\right). \end{aligned} \quad (15)$$

[6] These set of base-points can either be pre-assigned or varied on-the-fly depending on channel condition for better performance.

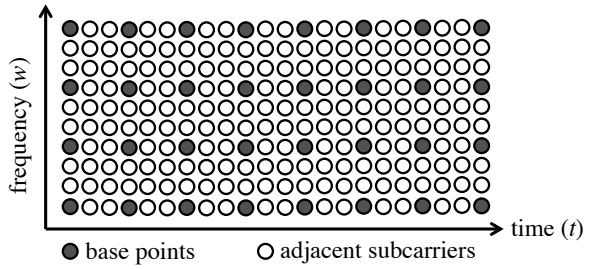


Fig. 2 Illustration of the frame structure of a wideband system. FADE only computes explicit matrix inverses at a small number of base points (in black); equalization at adjacent (in time and frequency) subcarriers is performed using one iteration of the implicit accelerated Neumann series recursion in (15).

By precomputing $\mathbf{A}_w^{-1}\mathbf{y}_{w+1}^{\mathrm{MF}}$, all subsequent operations in (15) consist of matrix-vector multiplications—this is one of the reasons why FADE requires low computational complexity. Another reason is the fact that for massive MU-MIMO only very few base points are required due to channel hardening [45]. Specifically, the entries in the Gram matrices $\mathbf{G}_w$ between adjacent subcarriers become similar in magnitude (or flat) as $B \to \infty$ [46].

Since FADE avoids computation of a matrix inverse $\mathbf{A}_{w'}^{-1}$ for adjacent subcarriers $w' \notin \Omega$, we compute the quantities $\mu^{(i)}$ and $\rho^{(i)}$ using the approximations outlined in Section 4.4 for implicit methods to compute approximate LLR values. Note that these approximations require the computation of $\mathbf{D}_w^{-1}$, which can be obtained efficiently from the squared column norms of the channel matrix $\mathbf{H}_w$.

Although FADE as discussed above only exploits FD correlation, it can be extended to exploit correlation in the time domain as well. For example, the inverse of the wth subcarrier of the $(t-1)$th symbol can be used as the initial estimate for the wth subcarrier of tth symbol. Figure 2 illustrates how FADE can exploit FD and TD correlations. In the following, we will show that FADE is not only computationally extremely efficient but also achieves near-exact performance, even for small BS-to-user-antenna ratios.

7 Computational Complexity

We now compare the computational complexity of existing and the proposed (approximate) inversion methods in terms of real-valued multiplications as an indication

Table 1 Complexity of different initialization methods.

$2\|\mathbf{A}^H\|_\infty^{-1}\mathbf{A}^H$	$(U+B)^{-1}\mathbf{I}_U$	$\mathbf{D}^{-1}$	Init. 1	Init. 2
$2U^2+1$	0	0	$2U^2+U+1$	$2U^2+U$

Table 2 Complexity of exact and approximate explicit matrix inversion methods for $K_{\max} \geq 1$.

Method	$\alpha\mathbf{D}^{-1}$	$\alpha\mathbf{A}^H$
Neumann recursion	$2BU^2+(K_{\max}-1)(2U^3)+(2U^2-2U)$	$2BU^2+(K_{\max}+1)(2U^3)$
Schultz recursion	$2BU^2+(K_{\max}-1)(6U^3)+(2U^2-2U)$	$2BU^2+K_{\max}4U^3$
3rd order recursion	$2BU^2+K_{\max}10U^3$	$2BU^2+K_{\max}6U^3$
Cholesky decomposition	$2BU^2+\frac{10}{3}U^3-\frac{4}{3}U$	

Table 3 Complexity of implicit matrix inversion methods.

Method	$\alpha\mathbf{D}^{-1}$ or $(U+B)^{-1}\mathbf{I}$	$2\|\mathbf{A}^H\|_\infty^{-1}\mathbf{A}^H$
Neumann	$2BU^2+K_{\max}(4U^2+2U)+2U$	$2BU^2+K_{\max}8U^2+4U^2$
CG	$2BU^2+(K_{\max}+1)(4U^2+20U)$	
GS	$2BU^2+K_{\max}4U^2+2U$	
Cholesky	$\frac{2}{3}U^3+2BU^2+4U^2-\frac{2}{3}U$	

of hardware efficiency.[7] For each complex-valued multiplication, we assume four real-valued multiplications and two real-valued additions. We also exploit symmetries (e.g., the fact that $\mathbf{A}$ is Hermitian) and avoid multiplications with zeros and ones.

7.1 Initialization Methods

Table 1 compares the complexity for all initialization methods discussed in Section 5.2 that can be implemented without an eigenvalue decomposition. Note that these complexity results ignore the computation of the regularized Gram matrix $\mathbf{A}$.

Evidently, computing $\mathbf{D}^{-1}$ as in [6] and $(U+B)^{-1}\mathbf{I}_U$ as in [9] does not require any multiplications. The complexity of the traditional initializer $2\|\mathbf{A}^H\|_\infty^{-1}\mathbf{A}^H$ is dominated by the term $\|\mathbf{A}^H\|_\infty^{-1}$ and requires a total of $2U^2+1$ real-valued multiplications. The complexity of the proposed initialization methods, Initialization 1 and 2, is $2U^2+U+1$ and $2U^2+U$, respectively. The complexity of both methods is dominated by the computation of the entry-wise norm of $\mathbf{D}^{-1}\mathbf{A}$. While the multiplication count for both initializers are very similar, Initialization 1 requires only one reciprocal operation whereas Initialization 2 requires U such operations. In what follows, we exclusively focus on Initialization 1, since Initialization 2 provides only slightly better performance (cf. Figure 1 and the discussion in Section 5.4).

[7] While the processing latency is another important design parameter in practice, it typically depends on (i) the data dependencies of the used algorithm, (ii) the hardware architecture (parallel or serial), and (iii) the computing fabric (e.g., GPU, FPGA, or ASIC). Hence, we limit our results on complexity aspects only—a detailed latency analysis would require hardware designs and is left for future work.

Table 4 Break-Even Point for Implicit Inversion. The break-even point is the smallest U such that the method exhibits lower complexity than the implicit Cholesky decomposition.

$K_{\max}$	0	1	2	3	4	5	6
$\mathbf{D}^{-1}$ and $(U+B)^{-1}\mathbf{I}$	1	3	8	14	19	25	31
$\alpha^{\text{opt}}\mathbf{D}^{-1}$	1	5	11	16	22	28	33
$2\|\mathbf{A}^H\|_\infty^{-1}\mathbf{A}^H$	1	16	28	40	52	64	76
CG	5	12	18	24	30	36	42
GS	1	3	7	13	19	25	31

7.2 Explicit Series Expansions and Exact Inversion

Table 2 compares the complexity of exact inversion via the Cholesky factorization and of various explicit series expansions as discussed in Section 3. All results in this table include the complexity of computing $\mathbf{A}$, which is necessary for all considered explicit methods. The complexity of Cholesky-based exact inversion scales with U^3 and is lower than the complexity of a standard[8] matrix multiplication. The complexity of the proposed explicit series expansions depends on two factors: the initialization matrix and the iteration count $K_{\max}$.

[8] More efficient matrix-multiplication algorithms, such as Strassen's algorithm which scales with $U^{2.8074}$, could be used [47]; the irregularity of such algorithms, however, renders efficient hardware designs difficult.

7.2.1 Impact of the Initialization Matrix

The initialization matrix $\alpha\mathbf{D}^{-1}$, for example, causes the intermediate terms $\tilde{\mathbf{A}}_K^{-1}\mathbf{A}$ to not be Hermitian, in general. In contrast, using $2\|\mathbf{A}^H\|_\infty^{-1}\mathbf{A}^H$ ensures that $\tilde{\mathbf{A}}_K^{-1}\mathbf{A}$ is Hermitian, leading to different operation counts for these two initialization matrices.

7.2.2 Impact of the Iteration Count

The case $K_{\max}=0$ corresponds to the computation of the initialization matrix as summarized in Table 1. For $K_{\max}=1$, the Neumann series expansion with the initialization matrix $\alpha_{\text{opt}}\mathbf{D}^{-1}$ leads to

$$\widetilde{\mathbf{A}}_1^{-1}=2\alpha_{\text{opt}}\mathbf{D}^{-1}-\alpha_{\text{opt}}^2\mathbf{D}^{-1}\mathbf{A}\mathbf{D}^{-1},$$

which requires only column and row scaling of $\mathbf{A}$. Hence, the associated complexity scales only in U^2. In this case, the truncated Neumann series approximation exhibits lower complexity than the explicit Cholesky-based inverse and is an attractive method for explicit equalization in massive MU-MIMO systems [6]. For $K_{\max}=1$ and the initialization matrix $2\|\mathbf{A}^H\|_\infty^{-1}\mathbf{A}^H$, however, the complexity of the truncated Neumann series expansion is larger than that of the explicit Cholesky-based inverse, because of the matrix multiplication required for the term $\tilde{\mathbf{A}}_0^{-1}\mathbf{A}$.

For $K_{\max}=1$ and $\alpha_{\text{opt}}\mathbf{D}^{-1}$, the Neumann recursion coincides to the Schultz recursion. For $K_{\max}>1$, however, the Schultz recursion requires two matrix-matrix multiplications per iteration, resulting in substantially higher complexity. Similarly, the 3$^{\text{rd}}$ order recursion requires three such operations per iteration. As a result, both of these explicit higher-order recursions are unattractive in terms of complexity despite the fact they enable fast convergence (cf. Section 5.4).

7.3 Implicit Series Expansions and Exact Inversion

Table 3 compares the complexity of various implicit equalization schemes, including Cholesky-based exact matrix inversion, various approximate series expansions, as well as iterative methods. As expected, the complexity of implicit methods is significantly lower than that of explicit methods (cf. Table 2). Furthermore, the complexity of Cholesky-based implicit equalization scales with U^3, whereas all other approximate methods scale only with U^2.

As for explicit, approximate methods, the complexity of the implicit Neumann recursion depends on the initialization matrix. The choice $\mathbf{A}_0^{-1}=\alpha\mathbf{D}^{-1}$ requires one matrix-vector multiplication per iteration; the choice $\mathbf{A}_0^{-1}=2\|\mathbf{A}^H\|_\infty^{-1}\mathbf{A}^H$ requires two such operations and, hence, is less attractive (also from a convergence point-of-view; see Section 5.4). Table 3 also includes the complexity of CG [10] and GS [10], which scales quadratically in U. As for the implicit Neumann recursion, GS must be initialized by computing $\tilde{\mathbf{y}}_0=\mathbf{D}^{-1}\mathbf{y}^{\text{MF}}$, whereas CG uses an all zero vector for the initialization vector. In addition, we see in Table 3 that the complexity of the exact Cholesky-based approach scales cubically in U.

We emphasize that the method that exhibits the lowest computational complexity is not immediately clear from Table 3. As it turns out, the implicit Cholesky decomposition often has the lowest complexity depending on U and $K_{\max}$, mainly due to the rather small constant of 2/3 in front of the U^3 term. Table 4 provides an overview of this rather surprising behavior by listing the break-even points, i.e., the smallest value of the number of users U such that the complexity of a given approximate implicit method is lower than that of the exact, implicit Cholesky decomposition. To ensure a fair comparison, we take into account the complexity required to compute the necessary initialization matrices (i.e., $\mathbf{D}^{-1}$, $\alpha^{\text{opt}}\mathbf{D}^{-1}$, and $2||\mathbf{A}^H||_\infty^{-1}\mathbf{A}^H$). We observe that the Neumann recursion, CG, and GS, are only competitive with the exact, implicit Cholesky decomposition for a very small number of iterations $K_{\max}$. In these cases, GS exhibits the lowest complexity among all considered implicit equalization methods.

7.4 Complexity of FADE

We now assess the complexity of the proposed frequency-adaptive equalizer (FADE). Since this method combines two methods: (i) an exact, explicit inversion using the Cholesky decomposition at each base point and (ii) an approximate, implicit Neumann recursion update on adjacent subcarriers (15), the total complexity of FADE is an average of the two methods.

The complexity of explicit inversion using the Cholesky decomposition is shown in Table 2; the complexity of the implicit Neumann recursion update in (15) is given by $8BU+8U^2+4U$. Let $p\in[0,1]$ and $\overline{p}=1-p\in[0,1]$ be the percentage of base-point subcarriers and the percentage of adjacent subcarriers, respectively. For example, $p=\frac{4}{10}$ and $\overline{p}=\frac{6}{10}$ in Figure 2. Then, the average number of real-valued multiplications required per subcarrier is simply the weighted sum of the two parts given by

$$p\big(2BU^2+\tfrac{10}{3}U^3-\tfrac{4}{3}U\big)+\overline{p}\,(8BU+8U^2+4U). \quad (16)$$

The parameter p controls a performance/complexity trade-off—large values of p perform more explicit matrix inversions, which result in high complexity but

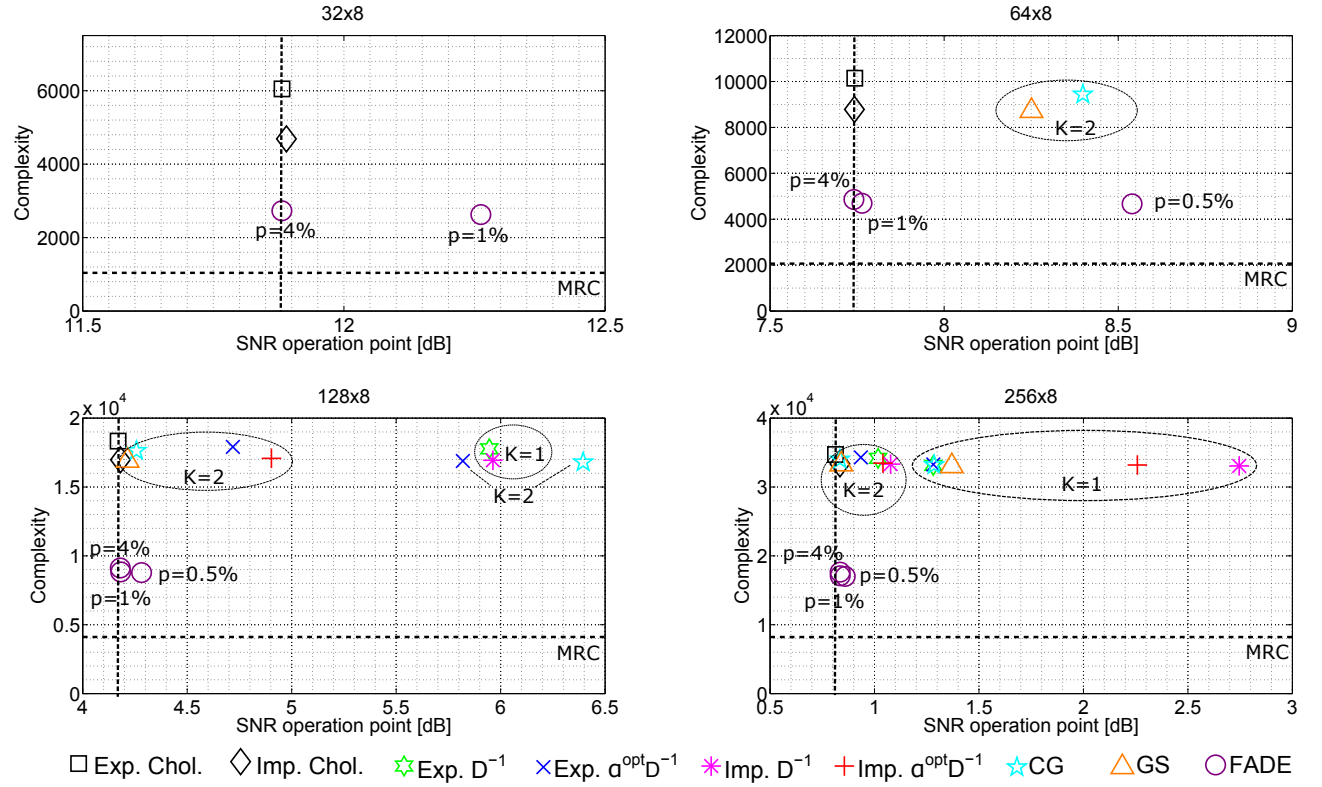


Fig. 3 Error-rate performance vs. complexity trade-off for different antenna configurations (we use the notation $B \times U$). The proposed FADE algorithm outperforms all considered exact/approximate methods for all antenna configurations and operates close to the complexity limit of maximum ratio combining (MRC). Furthermore, implicit methods generally outperform explicit methods in terms of complexity. The complexity is defined as the number of real-valued multiplications and the performance as the SNR operation point, which is the minimum SNR that is required to achieve a BLER of 10%.

deliver excellent error-rate performance; small values of p perform less explicit inversions which reduce the complexity at the cost of error-rate performance—this trade-off is studied next.

8 Performance/Complexity Trade-offs

We now investigate the performance/complexity trade-offs associated with the proposed equalizers and that of existing solutions.

8.1 Simulation Setup and Performance/Complexity Metrics

To evaluate the error-rate performance of the proposed soft-output data detectors, we consider a 3GPP-LTE uplink system [48] with $U = 8$ single-antenna user terminals and $B \in \{32, 64, 128, 256\}$ BS antennas. For all simulations, we consider a 20 MHz bandwidth with 1200 subcarriers, and we use 64-QAM with 3GPP turbo code of rate 3/4. To consider frequency and spatial correlation, we used a WINNER-Phase 2 channel model [49] with 8.9 cm antenna spacing; the maximum delay spread for this model is six taps. All simulations assume perfect synchronization and channel-state knowledge. To assess the error-rate performance without the need of many different performance curves, we use the so-called SNR operation point [50], which is defined as the minimum SNR required to achieve 10% block error-rate (BLER) for $U = 8$ users, which is representative for reliable transmission in LTE-based systems. The BLER is obtained via Monte-Carlo simulations averaged over 2000 transport blocks (TB), each consisting of 75,376 bits. To assess the computational complexity, we use the real-valued multiplication counts in Section 7. For the initialization matrix $\alpha^{\text{opt}}\mathbf{D}^{-1}$, we use $\alpha^{\text{opt}} = U\|\mathbf{D}^{-1}\mathbf{A}\|^{-2}$.

8.2 Trade-off Comparison

Figure 3 shows the performance/complexity trade-offs for all considered exact, approximate, explicit, and implicit methods, as well as FADE. First, we emphasize that the matched filter equalizer (which is equivalent to $K_{\max} = 0$) achieves the lowest complexity but is

unable to achieve 10% BLER for all considered antenna configurations. In contrast, the explicit matrix inversion using the Cholesky decomposition leads to exact MMSE equalization performance, while requiring the highest complexity. The implicit Cholesky decomposition achieves a BLER close to that of the exact MMSE detector for all the considered antenna configurations at slightly lower complexity—the performance loss of this implicit data detector comes from the SINR approximation in Section 4.4.

For $B = 32$, we see that only Cholesky-based exact inversion and FADE are able to achieve (near-optimal) performance. Furthermore, FADE with $p = 4\%$ exhibits the same performance at $2\times$ lower complexity. For $B = 64$, CG and GS are the only approximate, implicit methods that achieve 10% BLER; these methods, however, exhibit a similar complexity as the implicit Cholesky decomposition at about 0.5 dB performance loss. Again, FADE outperforms all methods in terms of performance and complexity. By increasing the number of BS antennas to $B \geq 128$, explicit as well as implicit Neumann series approximations start to approach the performance of the exact MMSE equalizer. However, only the implicit Neumann recursion enables lower complexity than the implicit Cholesky decomposition, which renders it attractive for massive MU-MIMO systems with large BS-to-user-antenna ratios. FADE reduces the complexity by more than $2\times$ at near-optimal performance for only 1% base points.

In summary, by exploiting frequency correlation, FADE is able to significantly reduce the complexity compared to all other methods, by eliminating the need to compute the regularized Gram matrix $\mathbf{A}$ at all subcarriers. We also observe that the complexity advantage of FADE becomes more pronounced for larger BS antenna arrays where (i) the complexity of computation of Gram matrix becomes the dominating operation and (ii) channel hardening enables us to use fewer base points [45]. We finally note that FADE can be combined with decentralized baseband architectures in which the antennas are distributed and baseband processing is performed on multiple computing fabrics in parallel in a decentralized fashion [51]; such systems have the potential to enable antenna arrays with thousands of BS antennas.

9 Conclusions

We have analyzed the performance and complexity of various exact, approximate, explicit, and implicit equalization schemes for wideband massive MU-MIMO systems that use SC-FDMA. Our results show that for small BS-to-user antenna ratios, exact and implicit Cholesky decomposition-based equalization methods achieve the best trade-off; for large BS-to-user antenna ratios—if the number of BS antennas is roughly $2\times$ larger than the number of user antennas—approximate and implicit methods such as conjugate gradients, Gauss-Seidel, or accelerated implicit Neumann series approximations in combination with our post-equalization SINR approximation enable further reductions in computational complexity at virtually no performance loss. Finally, we have shown that by combining the advantages of exact explicit and approximate implicit equalization using the proposed frequency adaptive equalizer (FADE), we can exploit frequency (and time) correlation in wideband massive MU-MIMO systems to achieve near-optimal error-rate performance at only 50% of the complexity of competitive methods. A hardware integration of the proposed algorithms (such as in [6,12,38,52]) on modern computing fabrics, and corresponding throughput and latency measurements are part of ongoing work.

Acknowledgements The work of MW, BY, KL and JRC was supported by Xilinx Inc. and by the US National Science Foundation (NSF) under grants CNS-1717218, ECCS-1408370, CNS-1265332, and ECCS-1232274. The work of CS was supported by Xilinx Inc. and by the US NSF under grants ECCS-1408006, CCF-1535897, CAREER CCF-1652065, and CNS-1717559.

A Proofs and Derivations

A.1 Proof of Lemma 1

We use [33, Thm. 4.20], which establishes that for a given matrix $\mathbf{P} \in \mathbb{C}^{U\times U}$ for which $\lim_{k\to\infty} \mathbf{P}^k = \mathbf{0}_{U\times U}$, we have $(\mathbf{I}_U - \mathbf{P})^{-1} = \sum_{k=0}^{\infty} \mathbf{P}^k$ and $\mathbf{I}_U - \mathbf{P}$ is invertible. As a consequence, by defining $\widetilde{\mathbf{A}}_0^{-1}\mathbf{A} = \mathbf{I}_U - \mathbf{P}$ and assuming that $\lim_{k\to\infty}(\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A})^k = \mathbf{0}_{U\times U}$, we have

$$\mathbf{A}^{-1}\widetilde{\mathbf{A}}_0 = (\widetilde{\mathbf{A}}_0^{-1}\mathbf{A})^{-1} = \textstyle\sum_{k=0}^{\infty}(\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A})^k,$$

which can be rewritten to the accelerated Neumann series in (4), since $\widetilde{\mathbf{A}}^{-1}$ was assumed to be full rank.

A.2 Proof of Lemma 4

We start by rewriting the residual error term as a function of $\mathbf{A}_0^{-1}$. We have the following identities:

$$\begin{aligned}\mathbf{e}_{K_{\max}} &= \tilde{\mathbf{y}}_{K_{\max}} - \mathbf{A}^{-1}\mathbf{y}^{\mathrm{MF}} = (\widetilde{\mathbf{A}}_{K_{\max}}^{-1} - \mathbf{A}^{-1})\mathbf{y}^{\mathrm{MF}} \\ &= \Big(-\textstyle\sum_{k=K_{\max}+1}^{\infty}(\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A})^k\widetilde{\mathbf{A}}_0^{-1}\Big)\mathbf{y}^{\mathrm{MF}} \\ &= -(\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A})^{K_{\max}+1}\mathbf{A}^{-1}\mathbf{y}^{\mathrm{MF}}.\end{aligned}$$

By using basic properties of induced norms, we get the following inequality:

$$\|\mathbf{e}_{K_{\max}}\| \leq \|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|^{K_{\max}+1}\|\tilde{\mathbf{y}}\|,$$

where we define $\tilde{\mathbf{y}} = \mathbf{A}^{-1}\mathbf{y}^{\mathrm{MF}}$.

A.3 Derivation of Initialization 1

We start by noting that squaring both sides in (5) results in the equivalent sufficient condition $\|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|^2 < 1$. Furthermore, by assuming the spectral norm (which is a consistent norm), we have $\|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|^2 \leq \|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|_F^2$, which enables us to obtain a more restrictive sufficient condition that allows the design of efficient initializers:

$$\|\mathbf{I}_U - \widetilde{\mathbf{A}}_0^{-1}\mathbf{A}\|_F^2 < 1. \tag{17}$$

The initialization method developed next[9] is of the form $\widetilde{\mathbf{A}}_0^{-1} = \mathbf{S}\mathbf{D}^{-1}$, where $\mathbf{S}$ is a diagonal scaling matrix that is designed to meet condition (17). Let $\mathbf{W}$ contain the diagonal part of $\mathbf{D}^{-1}\mathbf{A}$ and $\mathbf{Q}$ the off-diagonal part. We define

$$\begin{aligned} f &= \|\mathbf{I}_U - \mathbf{S}\mathbf{D}^{-1}\mathbf{A}\|_F^2 = \|\mathbf{I}_U - \mathbf{S}(\mathbf{W}+\mathbf{Q})\|_F^2 \\ &= \|\mathbf{I}_U - \mathbf{S}\mathbf{W}\|_F^2 + \|\mathbf{S}\mathbf{Q}\|_F^2, \end{aligned} \tag{18}$$

and seek a diagonal scaling matrix $\mathbf{S}$ that minimizes f.

We define the diagonal scaling matrix to have the form $\mathbf{S} = \alpha\,\mathbf{I}$, which leads to $f = \sum_{i=1}^{U} |1 - \alpha\, W_{i,i}|^2 + |\alpha|^2\|\mathbf{Q}\|_F^2$. We now find the optimum scaling parameter α^{opt} by computing $\partial f/\partial\alpha^* = 0$ and solving for α. Standard manipulations yield

$$\alpha^{\text{opt}} = \big(\textstyle\sum_{i=1}^{U} W_{i,i}^*\big)\|\mathbf{D}^{-1}\mathbf{A}\|_F^{-2},$$

where $W_{i,i}^*$ are the complex conjugates of the diagonal entries of $\mathbf{W}$. Since $\mathbf{W} = \mathbf{I}_U$, we get $\alpha^{\text{opt}} = U\|\mathbf{D}^{-1}\mathbf{A}\|_F^{-2}$. Consequently, the first initialization matrix is $\widetilde{\mathbf{A}}_0^{-1} = \alpha^{\text{opt}}\mathbf{D}^{-1}$.

A.4 Derivation of Initialization 2

For the second initialization method, we follow the derivation in Appendix A.3 but with a more general diagonal scaling matrix of the form $\mathbf{S} = \text{diag}\,(\alpha_1, \ldots, \alpha_U)$. We obtain

$$f = \textstyle\sum_{i=1}^{U} |1 - \alpha_i\, W_{i,i}|^2 + |\alpha_i|^2\|\mathbf{r}_i\|_2^2,$$

where $\mathbf{r}_i$ corresponds to the i^{th} row of $\mathbf{Q}$. To find the optimal scaling parameters α_i, $i = 1, \ldots, U$, we set $\partial f_i/\partial\alpha_i^* = 0$ and solve for α_i. Standard manipulations yield

$$\alpha_i^{\text{opt}} = W_{i,i}^*/(|W_{i,i}|^2 + \|\mathbf{r}_i\|_2^2),\ i = 1, \ldots, U,$$

and we use the fact that $W_{i,i} = 1$, $\forall i$. Consequently, the second initialization matrix is

$$\widetilde{\mathbf{A}}_0^{-1} = \text{diag}(\alpha_1^{\text{opt}}, \ldots, \alpha_U^{\text{opt}})\mathbf{D}^{-1}.$$

[9] This initialization scheme can also be used by replacing $\mathbf{D}^{-1}$ with an arbitrary matrix $\mathbf{X}$ that is close to the exact inverse $\mathbf{A}^{-1}$.

References

1. F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.
2. T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.
3. Y.-H. Nam, B. L. Ng, K. Sayana, Y. Li, J. Zhang, Y. Kim, and J. Lee, "Full-dimension MIMO (FD-MIMO) for next generation cellular technology," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 172–179, Jun. 2013.
4. H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving "massive MIMO" spectral efficiency with a not-so-large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3266–3239, Sept. 2012.
5. H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. on Commun.*, vol. 61, no. 4, pp. 1436–1449, April 2013.
6. M. Wu, B. Yin, G. Wang, C. Dick, J. R. Cavallaro, and C. Studer, "Large-scale MIMO detection for 3GPP LTE: Algorithm and FPGA implementation," *IEEE J. Sel. Topic of Sig. Proc.*, vol. 8, no. 5, pp. 916 – 929, Oct. 2014.
7. H. Prabhu, J. Rodrigues, O. Edfors, and F. Rusek, "Approximative matrix inverse computations for very-large MIMO and applications to linear pre-coding systems," in *Proc. IEEE WCNC*, April 2013, pp. 2710–2715.
8. L. Dai, X. Gao, X. Su, S. Han, I. Chih-Lin, and Z. Wang, "Low-complexity soft-output signal detection based on Gauss–Seidel method for uplink multiuser large-scale MIMO systems," *IEEE Trans. on Vehicular Techn.*, vol. 64, no. 10, pp. 4839–4845, Nov. 2015.
9. Z. Lu, J. Ning, Y. Zhang, T. Xie, and W. Shen, "Richardson method based linear precoding with low complexity for massive MIMO systems," in *Proc. IEEE VTC*, May 2015.
10. B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, "Conjugate gradient-based soft-output detection and precoding in massive MIMO systems," in *Proc. IEEE GLOBECOM*, Dec 2014, pp. 3696–3701.
11. H. Prabhu, O. Edfors, J. Rodrigues, L. Liu, and F. Rusek, "Hardware efficient approximative matrix inversion for linear pre-coding in massive MIMO," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, June 2014, pp. 1700–1703.
12. B. Yin, M. Wu, G. Wang, C. Dick, J. R. Cavallaro, and C. Studer, "A 3.8 Gb/s large-scale MIMO detector for 3GPP LTE-Advanced," in *Proc. IEEE ICASSP*, May 2014.
13. N. E. Tunali, M. Wu, C. Dick, and C. Studer, "Linear large-scale MIMO data detection for 5G multi-carrier waveform candidates," *49th Asilomar Conference on Signals, Systems, and Computers*, Nov. 2015.
14. A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. Bolcskei, "VLSI implementation of MIMO detection using the sphere decoding algorithm," *IEEE Journal of solid-state circuits*, vol. 40, no. 7, pp. 1566–1577, June 2005.
15. C. Studer, A. Burg, and H. Bölcskei, "Soft-output sphere decoding: algorithms and VLSI implementation," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 2, pp. 290–300, Sep. 2008.
16. K. Wong, C. Tsui, R. Cheng, and W. Mow, "A VLSI architecture of a K-best lattice decoding algorithm for MIMO channels," in *Proc. IEEE ISCAS*, vol. 3, May 2002, pp. 273–276.
17. S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, pp. 1941–1988, Nov. 2015.
18. M. Wu, C. Dick, J. R. Cavallaro, and C. Studer, "Iterative detection and decoding in 3GPP LTE-based massive MIMO systems," in *Proc. EUSIPCO*, Sept. 2014, pp. 96–100.

19. M. Cirkic and E. G. Larsson, "SUMIS: near-optimal soft-in soft-out MIMO detection with low and fixed complexity," *IEEE Trans. on Sig. Proc.*, vol. 62, no. 12, pp. 3084–3097, May 2014.
20. T. Datta, N. Ashok Kumar, A. Chockalingam, and B. Sundar Rajan, "A novel MCMC algorithm for near-optimal detection in large-scale uplink mulituser MIMO systems," in *Proc. IEEE ITA*, Feb. 2012, pp. 69 –77.
21. T. L. Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing (CHEMP) receiver in large-scale MIMO systems," *IEEE J. Sel. Topics of Sig. Proc.*, vol. 8, no. 5, pp. 847–860, Oct. 2014.
22. C. Jeon, R. Ghods, A. Maleki, and C. Studer, "Optimality of large MIMO detection via approximate message passing," in *Proc. IEEE ISIT*, June 2015, pp. 1227–1231.
23. J. Ketonen, J. Karjalainen, M. Juntti, and T. Hänninen, "MIMO detection in single carrier systems," in *Proc. EUSIPCO*, Aug. 2011, pp. 654–658.
24. G. Berardinelli, C. Manchon, L. Deneire, T. Sorensen, P. Mogensen, and K. Pajukoski, "Turbo receivers for single user MIMO LTE-A uplink," in *Proc. IEEE VTC*, Apr. 2009, pp. 1–5.
25. S. Okuyama, K. Takeda, and F. Adachi, "Iterative MMSE detection and interference cancellation for uplink SC-FDMA MIMO using HARQ," in *Proc. IEEE ICC*, June 2011, pp. 1–5.
26. C. Studer, S. Fateh, and D. Seethaler, "ASIC implementation of soft-input soft-output MIMO detection using MMSE parallel interference cancellation," *IEEE J. Solid-State Circuits*, vol. 46, no. 7, pp. 1754–1765, Jul. 2011.
27. A. Burg, S. Haene, D. Perels, P. Luethi, N. Felber, and W. Fichtner, "Algorithm and VLSI architecture for linear MMSE detection in MIMO-OFDM systems," in *Proc. IEEE ISCAS*, May 2006, pp. 4102–4105.
28. P. Luethi, A. Burg, S. Haene, D. Perels, N. Felber, and W. Fichtner, "VLSI implementation of a high-speed iterative sorted MMSE QR decomposition," in *Proc. IEEE ISCAS*, May 2007, pp. 1421–1424.
29. M. Karkooti, J. R. Cavallaro, and C. Dick, "FPGA implementation of matrix inversion using QRD-RLS algorithm," in *Proc. 44th Asilomar Conf. on Signals, Systems and Computers*, Nov. 2005, pp. 1625–1629.
30. S. Bellis, W. Marnane, and P. Fish, "Alternative systolic array for non-square-root Cholesky decomposition," in *IEEE Proc. on Computers and Digital Techniques*, vol. 144, no. 2, Mar 1997, pp. 57–64.
31. O. Maslennikow, V. Lepekha, A. Sergiyenko, A. Tomas, and R. Wyrzykowski, "Parallel implementation of Cholesky LL T-algorithm in FPGA-based processor," in *Parallel processing and applied mathematics*. Springer, Sept. 2007, pp. 137–147.
32. D. Yang, G. D. Peterson, H. Li, and J. Sun, "An FPGA implementation for solving least square problem," in *17th IEEE Symposium on Field Programmable Custom Computing Machines*, April 2009, pp. 303–306.
33. G. Stewart, *Matrix Algorithms: Basic decompositions*, 1998.
34. G. Schulz, "Iterative Berechung der reziproken Matrix," *Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 13, no. 1, pp. 57–59, 1933.
35. A. Burg, "VLSI circuits for MIMO communication systems," Ph.D. dissertation, ETH Zürich, 2006.
36. F. Soleymani, "A rapid numerical algorithm to compute matrix inversion," *International Journal of Mathematics and Mathematical Sciences*, 2012.
37. M. Altman, "An optimum cubically convergent iterative method of inverting a linear bounded operator in Hilbert space," *Pacific Journal of Mathematics*, vol. 10, no. 4, pp. 1107–1113, 1960.
38. M. Wu, C. Dick, J. R. Cavallaro, and C. Studer, "High-throughput data detection for massive MU-MIMO-OFDM using coordinate descent," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 63, no. 12, pp. 2357–2367, Dec. 2016.
39. A. Ben-Israel and D. Cohen, "On iterative computation of generalized inverses and associated projections," *SIAM Journal on Numerical Analysis*, vol. 3, no. 3, pp. 410–419, 1966.
40. D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
41. Y. Li, L. J. Cimini Jr, and N. R. Sollenberger, "Robust channel estimation for OFDM systems with rapid dispersive fading channels," *IEEE Trans. on Commun.*, vol. 46, no. 7, pp. 902–915, 1998.
42. S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing, 2009.
43. M. Borgmann and H. Bölcskei, "Interpolation-based efficient matrix inversion for MIMO-OFDM receivers," in *Proc. 38th Asilomar Conf. on Signals, Systems and Computers*, vol. 2, Nov. 2004, pp. 1941–1947.
44. D. Cescato and H. Bölcskei, "QR decomposition of Laurent polynomial matrices sampled on the unit circle," *IEEE Trans. on Info. Theory*, vol. 56, no. 9, pp. 4754–4761, Sept. 2010.
45. C. Jeon, Z. Li, and C. Studer, "Approximate Gram-matrix interpolation for wideband massive MU-MIMO systems," *arXiv preprint: 1610.00227*, Jun. 2017.
46. A. Farhang, N. Marchetti, L. E. Doyle, and B. Farhang-Boroujeny, "Filter bank multicarrier for massive MIMO," in *Proc. 80th IEEE Vehicular Technology Conf.*, Sept. 2014, pp. 1–7.
47. V. Strassen, "Gaussian elimination is not optimal," *Numerische Mathematik*, vol. 13, no. 4, pp. 354–356, 1969.
48. *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures (Release 10)*. 3GPP Organizational Partners TS 36.213 version 10.10.0, Jul. 2013.
49. L. Hentilä, P. Kyösti, M. Käske, M. Narandzic, and M. Alatossava, "Matlab implementation of the WINNER phase II channel model ver 1.1," Dec. 2007. [Online]. Available: https://www.ist-winner.org/phase_2_model.html
50. C. Studer and H. Bölcskei, "Soft–input soft–output single tree-search sphere decoding," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4827–4842, Oct. 2010.
51. K. Li, R. Sharan, Y. Chen, T. Goldstein, J. R. Cavallaro, and C. Studer, "Decentralized Baseband Processing for Massive MU-MIMO Systems," in *IEEE J. Emerging and Sel. Topics in Circuits and Systems (JETCAS)*, to appear in 2017.
52. K. Li, B. Yin, M. Wu, J. R. Cavallaro, and C. Studer, "Accelerating massive MIMO uplink detection on GPU for SDR systems," in *2015 IEEE Dallas Circuits and Systems Conference (DCAS)*, Oct 2015, pp. 1–4.