The limits of MIMO with Large Antenna Arrays

Pedro Valério Catarino Miguel
pedro.catarino.miguel@ist.utl.pt
Instituto Superior Técnico, Lisbon University, Portugal

Abstract—This paper engages different detection methods for MIMO systems, in an attempt to test and expand the limits of MIMO antenna arrays. One of the concerns of this work was to look for efficient detection algorithms that could allow decent results, while keeping a low complexity so that it could be possible to achieve large antenna arrays. With this in mind, this work presents the most prominent detection techniques for MIMO, starting with the linear receivers Zero-Forcing and MMSE, and then a Successive interference cancelation and lattice reduction-aided algorithms.

An alternative that requires channel state information at the receiver was studied, noting that this configuration allowed pre-processing to be done on the channel matrix, leading to advantageous communication conditions. Finally a study is made on the properties of large MIMO transmission channels, realizing several interesting facts that are a key motivation for the scalability of MIMO systems.

Keywords: MIMO detection, lattices, Zero-Forcing, MMSE, Successive interference cancelation, lattice reduction

I. INTRODUCTION

The field of information and communications technology (ICT) is facing a paradigm shift where portable devices and mobile Internet are steadily growing over personal computers and wired Internet services. Cellular technology started migrating towards data and Internet services with the introduction of third-generation (3G) and fourth-generation (4G) services, allowing new anytime/anywhere computing and multimedia applications that go from simple internet browsing up to mobile video streaming.

According to International Telecommunications Union (ITU), the mobile market will have over nine billion subscriptions in 2020, representing 113% of the projected world population, but with currently seven billion mobile-cellular telephone subscriptions it is now reaching a saturation point as it represents 95.5% of the world’s population. Even though the forecast points towards a stalling point, the telephone now has a surprisingly big penetration on the world’s population and has become a necessary tool for everyone’s life.

This combined with cellular applications that demand more and more bandwidth and data rates has led to a rapid growth on mobile data services. To demonstrate the quick evolution of the wireless technology in the recent past, we can observe the figure 2 and realize that in ten years there has been an expansion of

new radio equipment at the hardware platform level, going from 2G to 4G including WiFi, Bluetooth, open mobile handsets, software-defined radio, and most recently, open virtualized access points and base stations. At the radio physical layer, the cellular radio link speed has increased from about 2 Mb/s with early 3G systems in the year 2000 to 100 Mb/s with 4G (LTE and WiMax) systems using multiple-input–multiple-output (MIMO) radio technology which is the main focus in this work.

A key player in the networking industry [1] predicts that mobile generated traffic will exceed the traffic of fixed personal computers by 2015, as a result of the migration of most ICT services to mobile devices in the recent years.

With all these trends, large-scale delivery of internet applications on mobile devices will require faster radio access bit rates, improved spectrum efficiency, higher access system capacity and improved security on the open radio medium, among other needs. Most of the research and development on the wireless field today is focused on addressing these aspects, and this paper will briefly discuss some of those.

One of the problems on the migration into mobility is that mobile phones have limited batteries, which forces all the wireless communications technologies to be computationally efficient in order to minimize the power consumption. Another challenge is the fact that cellular networks keep getting populated with all the emerging smartphones, burdening the networks with heavy and complex data resulting in the need to distribute bandwidth efficiently.

MIMO is a technique that can tackle the problems stated before, and consists in multiple antennas at the transmitter and multiple antennas at the receiver. These multiple antennas at either the transmitter or receiver can be used to achieve diversity, an array power gain, resulting in performance and spectral efficiency gains.

If the MIMO system has access to the channel gain matrix, i.e. through a pilot test, independent signalling paths can be obtained by exploiting the structure of the channel matrix, allowing spatial multiplexing where multiple parallel streams of data are transmitted.

MIMO can then be divided into three categories; precoding, spatial multiplexing and diversity coding.

Precoding in more general terms is all the spatial processing that occurs at the transmitter, and can be looked as multi-stream beam forming. In single stream beam forming, each transmit antenna emits the same signal with appropriate phase and gain weighting such that the signal power is maximized at the receiver input. If the signals emitted from different antennas add up constructively, the multipath fading effect is reduced leading to an increased received signal gain. In line-of-sight propagation, beam forming results in a well-defined directional pattern. However, conventional beams are not a good analogy in cellular networks, which are mainly characterized by multipath...
propagation. When the receiver has multiple antennas, the transmit beam forming cannot simultaneously maximize the signal level at all of the receive antennas, and precoding with multiple streams is often beneficial. However, precoding requires knowledge of channel state information (CSI) at the transmitter and the receiver, which might limit its usability.

Spatial multiplexing requires a multiple-antenna setup. In spatial multiplexing, a high-rate signal is split into multiple lower-rate streams and each stream is transmitted from a different transmit antenna in the same frequency. If these signals arrive at the receiver antenna array with sufficiently different spatial signatures and the receiver has accurate CSI, it can separate these streams into (almost) parallel channels. Spatial multiplexing is a very powerful technique for increasing channel capacity at higher signal-to-noise ratios (SNR). The maximum number of spatial streams is limited by the smaller of the number of antennas at either the transmitter or the receiver. Spatial multiplexing can be used without CSI at the transmitter, but can be combined with precoding if CSI is available. Spatial multiplexing can also be used for simultaneous transmission to multiple receivers, known as space-division multiple access or multi-user MIMO, or also as the broadcast channel (in the information theory community), in which case CSI is required at the transmitter.

Diversity coding techniques are used when there is no channel knowledge at the transmitter. In diversity methods, a single stream (unlike multiple streams in spatial multiplexing) is transmitted, but the signal is coded using techniques called space-time coding. The signal is emitted from each of the transmit antennas with full or near orthogonal coding. Diversity coding exploits the independent fading in the multiple antenna links to enhance signal diversity. Because there is no channel knowledge, there is no beam forming or array gain from diversity coding. Diversity coding can be combined with spatial multiplexing when some channel knowledge is available at the transmitter.

II. MIMO DETECTION

The main idea behind spatial multiplexing is that signals sampled in the spatial domain at both ends are combined in order to create multiple parallel spatial data pipes (SISO channels), increasing the data rate, while possibly adding diversity to improve the quality of the communications by reducing the bit-error rate.

Spatial multiplexing in the context of MIMO has enabled unprecedented spectral efficiencies in wireless fading channels achieving high date-rates. However this gain of performance comes at the price of increased complexity in the detection of the receivers. This complexity presents a great challenge to the decoder’s implementation and is usually where most of the research is available and rising nowadays. In this chapter a background study will be done on some of the receivers that have been implemented so far. The performance of linear receivers will be analysed, as well as the successive interference cancellation (SIC) receivers. A lattice approach will be described in an attempt to expand the perception of the detection problem and to obtain better results.

One of the first examples of practical application of MIMO is the patent of Paulraj and Kailath [2] which introduced a technique for increasing the capacity of a wireless link using multiple antennas at both ends for application to broadcast digital TV.

The MIMO system that will be studied in this chapter is a single user point-to-point communication MIMO. This system is based on multiple antenna scenarios where both the transmitter and the receiver use several antennas, each one with separate radio frequency modules, and where the interfering channels are the radio links between all pair of transmit and receive antennas.

In this MIMO communications system with transmit antennas and receive antennas (with so that the linear system it gives rise to is determined) the relation between the transmitted and received signals can be modelled in the baseband as (1)

\[
y = Hx + n
\]

where \( y = [y_1, ..., y_{N_R}]^T \in C^{N_R \times 1} \) is the received signals vector and \( x = [x_1, ..., x_{N_T}]^T \in C^{N_T \times 1} \) is the transmitted signals vector. The radio links between each pair of transmit and receive antennas are represented by the channel matrix \( H \in C^{N_R \times N_T} \) in which its entries \( h_{ij} \) represent the complex coefficient associated with the link between the pair of a \( i^{th} \) receive antenna and the \( j^{th} \) transmit antenna. Each \( h_{ij} \) is taken from a zero-mean circularly symmetric complex Gaussian distribution with unit variance, which corresponds to having a variance equal to 1/2 in both real and imaginary components. In order to have an independent and identically distributed Rayleigh fading channel model the phase of each entry \( h_{ij} \) is uniformly distributed in \([0, 2\pi]\) and their amplitude has a Rayleigh distribution. In this model the vector \( n = [n_1, ..., n_{N_R}]^T \in C^{N_R \times 1} \) represents the noise vector that is added to the incoming signal vector. The entries of \( n \) are random variables taken from an independent circularly symmetric complex Gaussian with zero average and variance \( \sigma_n^2 \), so that both its real and imaginary components have variance \( \sigma_n^2/2 \). This noise model is usually called as zero-mean spatially white (ZMSW) noise [3].

In this work square quadrature amplitude modulation (QAM) constellations are used. The simulations are conducted with \( M \)-QAM constellations with \( M = 4, 16, 64 \) and the input symbols in each transmit antenna are taken from a finite complex constellation \( C \) constructed from the Cartesian product \( C = C_R \times C_R \), where \( C_R \) is the real alphabet

\[
C = \{-(\sqrt{M} - 1), ..., -3, -1, 1, +3, ..., + (\sqrt{M} - 1)\}. \tag{2}
\]

Although QAM is assumed, QPSK (quadrature phase shift keying) can also be used in MIMO, leading to the same results as 4-QAM as the radio waves are modulated the same way. In addition to that, most literature on MIMO spatial multiplexing uses QAM, and it is easier to relate the QAM constellations into lattices.

The average energy of the complex symbol taken from \( C \) is given by

\[
E_s = \frac{1}{M} \sum_{x_{ij} \in C} |x_{ij}|^2 \tag{3}
\]
Assuming, without loss of generality, that the filters at the receiver have impulse response \( h(t) \) normalised to \( \int |h(t)|^2 \, dt = 1 \).

Considering that each \( y_i \) receives the sum of \( N_r \) symbols weighted by unit power random variables, i.e., \( E[|h_{ij}|^2] = 1 \), on average it is valid to calculate the SNR at the receiver as
\[
\frac{E[||y||^2]}{E[||n||^2]} = \frac{E[\sum_{i=1}^{N_r} \sum_{j=1}^{N_t} |h_{ij}|^2]}{E[\sum_{i=1}^{N_r} n_i^2]}
\]
\[
= \frac{N_r \sigma_n^2}{N_t \sigma^2} = \frac{N_r \sigma_n^2}{\sigma^2}.
\]

Table 1 lists the values of the average energy \( E_s \) for the M-QAM modulations implemented in this work.

<table>
<thead>
<tr>
<th>Modulation</th>
<th>4-QAM</th>
<th>16-QAM</th>
<th>64-QAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_s )</td>
<td>2</td>
<td>10</td>
<td>42</td>
</tr>
</tbody>
</table>

In the following sections a series of receivers will be analysed, starting with the linear receivers such as zero-forcing (ZF) and minimum mean square error (MMSE); followed by the ordered successive interference cancelation (OSIC) algorithm, and finally the receivers using lattice reduction-aided (LRA).

One shall keep in mind that throughout this chapter, channel knowledge at the receiver is required in the model assumed, previously named as channel state information at the receiver (CSIR). Acquiring the channel knowledge is not an easy task, especially in fast fading channels, but this is not the focus of this thesis.

Throughout this work the performance of a receiver will be noted by plotting the symbol error rate (SER) as a function of the SNR. The diversity gain, or slope \( d \) is the metric used to evaluate the performance of the receivers. All receivers here analysed are well calibrated and the results are on pair with the results available in the literature.

### a) Linear receivers

Linear receivers consist of applying a linear transformation to the received vector followed by a quantization to the symbol alphabet, also known as slicing. They are the simplest of all receivers but also the ones with worse results. However, due to the low complexity, they are scalable to a fair number of antennas at the terminals as it will be studied on the fifth chapter.

One can estimate \( \hat{x} \) from
\[
x_{ZF} = H^{-1}y = H^{-1}Hx + H^{-1}n = x + H^{-1}n
\]
\[
\hat{x}_{ZF} = Q_c[x_{ZF}]
\]
which is commonly known as linear zero-forcing (ZF) receiver, as the interference caused by \( H \) is forced to be zero. In this thesis the channel matrix \( H \) was defined in the complex domain so in the equation (5), \( H^{-1} \) corresponds to the pseudo-inverse matrix, also known as Moore-Penrose matrix \( H^+ \).

\[
H^+ = (H^H H)^{-1} H^H
\]

Superscript \( (\cdot)^H \) denotes Hermitian operator which is a conjugation followed by transposition or vice-versa. One shall keep in mind that only matrices with non-zero determinants are invertible, and so it is required that \( N_R \geq N_T \). Note that the noise is enhanced by the \( H^+ \) transformation.

\[
\hat{x}_{ZF} = Q_c[H^+ y] = Q_c[H^+ x + H^+ n] = Q_c[x + H^+ n]
\]

The detected vector \( \hat{x}_{ZF} \), as obtained from (4.8), is in fact the solution to:

\[
\hat{x}_{ZF} = \arg \min_{x \in \mathbb{C}^N} ||y - Hx||^2
\]

This receiver solves the CVP by relaxing it to a search in a continuous neighbourhood instead of computing the distance between the received vector and every point in the lattice. It can also be seen as a linear transformation of the Voronoi regions of the \( Z^N \) by \( H \). The resulting regions are called ZF decision regions and correspond to the space where a lattice point will be interpreted as being closer to the lattice point associated with that region [4]. We can minimize the noise enhancement with minimum mean-square error (MMSE) receiver. In this receiver both the interference and the noise are considered in order to minimize the expected error. The main difference between these two linear receivers is that MMSE looks for a linear transformation \( W_{MMSE} \) that minimizes the mean square error between the estimated vector and the original vector.

\[
W_{MMSE} = \arg \min_W E[||Wy - x||^2]
\]

That can also be represented as:

\[
W_{MMSE} = \sigma_n^2 I_N + H^H (\sigma_n^2 I_N + H \cdot \sigma_n^2 I_N \cdot H^H)^{-1} H^H
\]

\[
\hat{x}_{MMSE} \text{ can be obtained by a linear transformation } W_{MMSE} \text{ followed by a quantization step that is the same as in ZF.}
\]

\[
\hat{x}_{MMSE} = Q_c[W_{MMSE} y]
\]
\[
= Q_c[W_{MMSE} Hx + W_{MMSE} n]
\]
\[
= Q_c[x + W_{MMSE} n]
\]

It is expected for MMSE to perform better than ZF because it solves CVP problem by relaxing the search in the continuous space where \( x \in \mathbb{C}^N \) but also for introducing a term that penalises large \( ||x|| \) and is proportional to the energy of the noise [4].

### b) Maximum-Likelihood Detection

Before posting the results of the linear receivers, one should establish the standard comparison of all the detectors relating to the MLD curve, the best curve possibly obtained, although mainly computational as its performance is extremely expensive computationally.
The maximum likelihood estimate \( \hat{x}_{\text{ML}} \) for \( x \) given \( y \) is

\[
\hat{x}_{\text{ML}} = \arg \max_{x \in \mathbb{C}} p(y|H,x)
\]  

(13)

Where assuming the ZMSW noise model, the probability density function of \( y \) given \( H \) and \( x \) can be written as

\[
p(y|H,x) = \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left(-\frac{||y - Hx||^2}{2\sigma_n^2}\right)
\]

Leading to

\[
\hat{x}_{\text{ML}} = \arg \max_{x \in \mathbb{C}} \frac{1}{(2\pi\sigma_n^2)^{N/2}} \exp\left(-\frac{||y - Hx||^2}{2\sigma_n^2}\right)
\]

\[
\hat{x}_{\text{ML}} = \arg \min_{x \in \mathbb{C}} ||y - Hx||^2
\]

(15)

The exponential parcel induces a growth in the search space for \( M \)-QAM constellations thus discouraging the use of brute force maximum-likelihood detection when \( N \) increases.

The figure 1 shows the performance of ZF, MMSE and ML detectors for 2x2 MIMO configuration. One can observe several happenings. The ML curve achieves diversity order equal to \( N \) as stated in the previous chapters, and so it serves as comparison because it achieves the best possible performance. Note that this simulation was run with few antennas as ML is a heavy detection algorithm, hence the 2x2 system. Looking now at ZF and MMSE curves, it is possible to observe that MMSE performs slightly better than ZF, because it solves the CVP problem by relaxing the search in the continuous space but also for introducing a term that penalises large \( ||x|| \) and is proportional to the energy of the noise. Regarding the inclination of the curve, both ZF and MMSE maintain a steady -1 slope which corresponds to a decrease of SER by a factor of 10 for every 10dB of SNR.

\[
\hat{x} = \arg \min \|QH-y-Rx\|^2
\]

(16)

Thanks to the upper triangular structure of \( R \), the detector performs a layered-search from the \( N_r \)th layer to the 1\textsuperscript{st} layer to find optimal solutions

One problem of this strategy is that if one symbol is incorrectly detected, all the following symbols are likely to be incorrect as well.

Instead of ZF equalization it is also possible perform linear MMSE equalization to obtain a MMSE version of the OSIC receiver, by adjusting the equation (16)

The performance of the OSIC receiver described above can be seen in figure 2. The OSIC based on ZF and MMSE is denoted as OSIC-ZF and OSIC-MMSE. One can see that these linear OSIC receivers still obtain curves with slope -1 as OSIC does not improve diversity but a large power gain.

This gain can be observed by comparing figure 2 with the results obtained in 1 however the simulations on figure 1 was run with a 2x2 MIMO system and here we are analysing a 3x3 MIMO configuration, so the gain is higher than if comparing under the same circumstances.
The pseudo-algorithm used to simulate LRA receivers is shown in Table 2.

Table 2: Pseudo-code of LRA detection.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shift and scale the constellation to have zero as a lattice point: ( y_{\text{red}} = \frac{1}{2}(y + Hp) )</td>
</tr>
<tr>
<td>2</td>
<td>Reduce the lattice basis ( H ) using CLLL: ( [H_{\text{red}}, M] = \text{CLLL}(H) )</td>
</tr>
<tr>
<td>3</td>
<td>Apply some detector to the CVP defined by ( (H_{\text{red}}, y_{\text{red}}) )</td>
</tr>
<tr>
<td>4</td>
<td>( z = \text{Detect}(H_{\text{red}}, y_{\text{red}}) )</td>
</tr>
<tr>
<td>5</td>
<td>Using ( z ) and ( M ) estimate the symbol in the original coordinate system ( \hat{x} = 2MQ_2[z] - p )</td>
</tr>
</tbody>
</table>

The performance of LRA receivers is now presented. A first analysis was performed on the linear receivers with lattice reduction which can be seen in Figure 3. Comparing the results of LRA-ZF and LRA-MMSE with the results from the last section OSIC-ZF and OSIC-MMSE, one can conclude that the LRA allows a desirable and visible gain. The second analysis of LRA receivers was performed on the linear receivers, but this time with OSIC schemes, leading to Figure 4. Comparing the results of these two figures with Figure 3, one can observe that the introduction of the OSIC approach allows a slight gain, however it is much less noticeable than the gain obtained by the increment of the LRA technique in the first two figures.

Summarising the information introduced hitherto, one should note that LRA-OSIC-MMSE receiver is the one that allows the best performance so far. However, an experimental approach will be made upon LRA-OSIC detection algorithms, which consists of doing a randomized detection of the OSIC algorithm, in an attempt of achieving results beyond LRA-OSIC receivers.

The third-order dispersion effects do not affect the form or the amplitude of the impulse. However, have a considerable effect on the propagation time, delaying it.
While lattice reduction-aided decoding has a relatively small complexity, its performance exhibits a widening gap to the maximum likelihood curve as the dimension increases. In order to improve its performance, this thesis presents an improved algorithm that can shorten that gap to ML detection, which is a randomized version of SIC, first proposed by Liu et al. (2011), and based on the work by Klein (2000).

Previously, an OSIC algorithm was described where a standard rounding to the nearest Gaussian integers was applied. However, in this algorithm, the standard rounding will be replaced by Klein’s randomized rounding, and a few candidate lattice points are sampled from a Gaussian-like distribution over the lattice. To find the closest lattice point, Klein’s algorithm is used to sample some lattice points and the closest among those samples is chosen. Lattice reduction increases the probability of finding the closest lattice point, and only needs to be run once during pre-processing.

One should calculate

\[ \hat{x}_i = \text{randRound} \left( \frac{\hat{y}_i - \sum \rho q_j R_j \hat{x}_j}{r_{ti}} \right). \tag{22} \]

where function \( \text{randRound} \) rounds the real and imaginary parts of \( r \) (denoted as \( r_{re} \) and \( r_{im} \)) to integers \( q_{re} \) and \( q_{im} \), respectively, according to the discrete Gaussian distribution defined as

\[ P(Q = q_{re}) = \frac{e^{-c(r_{re} - q_{re})}}{\sum_{q=-\infty}^{\infty} e^{-c(|r_{re} - q_{re}|)}} \tag{23} \]

For the real part and

\[ P(Q = q_{im}) = \frac{e^{-c(r_{im} - q_{im})}}{\sum_{q=-\infty}^{\infty} e^{-c(|r_{im} - q_{im}|)}} \tag{24} \]

for the imaginary part.

The parameter \( c_i \) is computed as

\[ c_i = \frac{\log(\rho)}{\min_i (r_{re})} r_{2i}^2 \tag{25} \]

with \( \rho \) being another parameter whose optimum value can be obtained from

\[ K = \left( \frac{\rho}{r_{2i}} \right)^2 \tag{26} \]

The number of candidate lattice points that are considered in the algorithm is represented by \( K \). The candidate list is then built by repeating \( K \) times the procedure for computing a lattice point with Equation (22). If the transmission is uncoded the final estimate will correspond to the closest of the \( K \) lattice point candidates. Due to the randomized nature of the algorithm, to avoid the possibility of the final estimate being further away than the one produced by the OSIC decoder, one of the \( K \) candidates should be obtained through roundoff and then scaling back the obtained solution into the original lattice.

Figure 5 depicts the performance of randomized decoding in comparison to the traditional ZF inverter and the standard OSIC detector. A 3x3 MIMO system was considered with 4 QAM modulation.

One may observe how the randomised version of OSIC captures the full diversity available, in contrast to traditional deterministic OSIC based on deterministic rounding at each iteration. The price to pay is the need to generate \( K \) candidates to select from, basically multiplying the number of operation by that factor \( K \).

III. SCALING UP MIMO

Previously it was shown that the more antennas the transmitter/receiver is equipped with and the more degrees of freedom the propagation channel can provide, the better the performance in terms of data rate or link reliability. There are a few wireless broadband standards based on this knowledge, like the LTE standard that allows for up to 8 antenna ports at the base station to take advantage of all these features. On a channel that varies rapidly as a function of time and frequency, and where circumstances permit coding across many channel coherence intervals, the achievable rate scales with min (nt, nr) log (1+SNR). The
gains in multiuser systems are even more impressive, because such systems offer the possibility to transmit simultaneously to several users and the flexibility to select what users to schedule for reception at any given point in time [9]. However, the price to pay for obtaining these new data rates with MIMO is increased complexity of the hardware and energy consumption of the signal processing at both ends. In point-to-point communications the complexity at the receiver requires more attention than the complexity at the transmitter because, for instance, the complexity of optimal signal detection grows exponentially with NT [10] [11]. On the other hand, in multiuser systems the greatest concern is on the complexity at the transmitter, as the coding schemes used to transmit information simultaneously to more than one user while maintaining decent levels of inter-user interference are rather complicated.

A decade ago, the large antenna arrays regime, when NT and NR increases, has not got beyond academic study, in that some asymptotic capacity scaling laws are known for ideal situations. However, this view is changing due to a few important system aspects in the large terminals regime have been discovered. For example, [12] showed that asymptotically as NT → ∞, and under realistic assumptions on the propagation channel, with a bandwidth of 20 MHz, a time-division multiplexing cellular system may accommodate more than 40 single-antenna users that are offered a net average throughput of 17 Mbits per second both in the reverse (uplink) and the forward (downlink) links, and a throughput of 3.6 Mbits per second with 95% probability! These rates are achievable without cooperation among the base stations and by relatively rudimentary techniques for CSI acquisition based on uplink pilot measurements.

When talking about very large MIMO, one thinks of systems that use antenna arrays with hundreds or more antennas. Very large MIMO entails a considerable number of antennas simultaneously serving a much smaller number of terminals. The disparity in number emerges as a desirable operating condition and a practical one as well. The number of terminals that can be simultaneously served is limited by the inability to acquire channel-state information for an unlimited number of terminals. Larger numbers of terminals can always be accommodated by combining very large MIMO technology with conventional time and frequency division multiplexing via OFDM. It is expected that in very large MIMO systems, each antenna unit uses extremely low power, in the order of nW. Ideally, the total transmitted power should be kept constant as one increases NT, meaning that the power per antenna should be proportional to 1/NT. It could also be possible to reduce the total transmitted power, but the need for multi-user multiplexing gains, errors in CSI, and interference would prevent these power savings in practice. Even considering these problems, the possibility of saving an order of magnitude in transmit power is desirable because it would open the possibility of achieving a better system performance under the same power constraints.

Another concern when studying very-large MIMO is the energy consumption of the cellular base stations. However, very-large MIMO designs can be made extremely robust to the failure of one or a few of the antenna units, meaning that a punctual failure would not harshly affect the system. A malfunctioning individual antenna can be swapped, in contrast to classical array designs, which use few antennas fed from a high-power amplifier.

Although the main advantages of large MIMO systems stand upon the increase of the diversity gain and data rate, there are also other advantages given by large dimension systems that cannot be seen in smaller systems, and are a key factor for the existence of large MIMO arrays. When the number of antennas at the terminals increases, the matrix H becomes larger, and the distribution of the singular values becomes less reliant on the entries of the channel matrix as long as they stay independent and identically distributed (iid) random variables.

This is a known result of the Marcenko-Pastur law [13] and states that if the entries of an nr × nt matrix H are zero mean iid with variance 1/nr, then the empirical distribution of the eigenvalues of H'H converges almost surely, as NT, nr →∞ with NT/NR → β, to the density function. A direct result of the Marcenko-Pastur law is that very tall or very wide channel matrices tend to be very well conditioned, meaning that the eigenvalue histogram of a single realisation converges to the average asymptotic eigenvalue distribution, resulting in a more deterministic channel as the channel dimension increases. This is usually called “channel hardening” in the literature [14], and its behaviour at large dimensions can be observed in the figure 6, where H'H was plotted with NT=NR=30 and 100, the entries of H are iid Gaussian entries with zero mean and unit variance.

The main property to retain is that as H grows in size, the diagonal terms of H'H become increasingly larger than non-diagonal terms, resulting in a near-diagonal matrix.
The discovery of channel hardening enabled some advantages in signal processing, more specifically in large dimension MIMO. One of these advantages rely on the fact that the inversion of large random matrices can be efficiently approximated with series expansion like Newman Series \[15\] and deterministic approximations. This is helpful because the linear detectors studied in this thesis like Zero-Forcing and minimum mean square error detectors perform matrix inversions in their algorithmic process, thus making them unviable detectors for very large systems, due to the excessive computation requirement of inverting for example 100x100 matrices.

However, due to its simplicity, we can still obtain decent results for large antenna arrays using these two simple detectors. Channel hardening turns these low complexity detection algorithms suited for large channels.

Figure 7 shows the performance of ZF and MMSE receivers but this time in large MIMO configurations. This time the ML curve was dismissed as it is a really heavy decoder, making it not feasible for comparison. The modulation chosen for this test was also 4-QAM as it’s the cheapest modulation, and this test is mainly focusing on hardware performance, hence the focus on cheap algorithms in order to reach large dimensions.

However, one can note that even at a high antenna regime, the linear detectors can maintain a steady slope. This is on pair with the previous results and the literature that states that the linear receivers can only maintain a slope of -1. As expected, increasing the number of antenna terminals does not influence the curve slope but rather the SER, that is slightly degraded by the increasing dimensions.

Another direct consequence from channel hardening and that could already be observed from the figure 6 is that when dimensions grow big, the Gram matrix tends to be near diagonal. By plotting the Frobenious distance of a Gram matrix to the correspondent Identity matrix \[\|\mathbf{H}^\dagger\mathbf{H} - \mathbf{I}\|_{Frob}\] we got the following figure. The script created random Channel matrices from nt=1 up to 400.

Observing this figure 8 we can note that as the number of antennas increases, the distance to the Identity matrix gets smaller, limited by an asymptotic zero. This similarity to diagonal matrices also allows us to obtain some results by applying the standard SVD precoding to large channels. And again, the results are feasible due to channel hardening.

Note that even though the number of antennas more than doubles, the performance of the detector does not alter much, meaning that it is a good and scalable algorithm.
IV. CONCLUSIONS

One of the objectives of this paper was to exploit certain channel characteristics that would make MIMO systems with large antenna arrays possible. Tests were run on popular receivers like ZF, MF and some non-linear variations with OSIC and LRA. Although the receivers that achieve the best performance are LRA-OSIC-MMSE and a randomized version of it, these algorithms ended up being computationally heavy as the number of antennas at the terminals increases, meaning that those were not scalable beyond 10 antennas.

In order to obtain large antenna arrays working for MIMO systems, we had to bet on cheaper algorithms like plain ZF and MMSE, and more ingenious detection methods involving channel information feedback that would allow the processing of the channel information at the receiver. By having access to the channel at the transmitter, it would then be possible to take advantage of the channel properties, creating individual parallel data pipes that are directly related to the channel matrix Eigen values. This channel simplification enables the use of large antenna arrays on MIMO systems, even though channel state information at the transmitter is required.

Throughout this work, some opportunities were identified as possible improvements for further development. The Newman series expansion was pointed as a way of decreasing the complexity of the inversion of matrices through a numeric approximation. This approximation may allow the inversion of large matrices in a feasible time, possibly enabling the use of linear receivers in very large antenna arrays.

Another possibility for achieving very large MIMO systems is to exploit the randomized decoder properties. This algorithm achieves full diversity at little complexity even for large MIMO configurations, so it might possible to take advantage of this fact and increase the dimension of the arrays used in MIMO systems.

7. REFERENCES


[8] Shuiyin Liu, Cong Ling, and Damien Stehlé, “Randomized lattice decoding,” IEEE Translations on information Theory, September 2011


