Fifty Years of MIMO Detection: The Road to Large-Scale MIMOs

Shaoshi Yang, Lajos Hanzo

Institutions: University of Southampton

Published on: 07 Sep 2015 - IEEE Communications Surveys and Tutorials (IEEE)

Topics: 3G MIMO, Multi-user MIMO, Spatial multiplexing, MIMO-OFDM and MIMO

Related papers:
- Scaling Up MIMO: Opportunities and Challenges with Very Large Arrays
- Massive MIMO for next generation wireless systems
- Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas
- An Overview of Massive MIMO: Benefits and Challenges
- Energy and Spectral Efficiency of Very Large Multiuser MIMO Systems
Fifty Years of MIMO Detection: The Road to Large-Scale MIMOs

Shaoshi Yang, Member, IEEE and Lajos Hanzo, Fellow, IEEE

Abstract—The emerging massive/large-scale multiple-input multiple-output (LS-MIMO) systems that rely on very large antenna arrays have become a hot topic of wireless communications. Compared to multi-antenna aided systems being built at the time of writing, such as the long-term evolution (LTE) based fourth generation (4G) mobile communication system which allows for up to eight antenna elements at the base station (BS), the LS-MIMO system entails an unprecedented number of antennas, say 100 or more, at the BS. The huge leap in the number of BS antennas opens the door to a new research field in communication theory, propagation and electronics, where random matrix theory begins to play a dominant role. Interestingly, LS-MIMOs also constitute a perfect example of one of the key philosophical principles of the Hegelian Dialectics, namely that “quantitative change leads to qualitative change”.

In this treatise, we provide a recital on the historic heritages and novel challenges facing LS-MIMOs from a detection perspective. Firstly, we highlight the fundamentals of MIMO detection, including the nature of co-channel interference (CCI), the generality of the MIMO detection problem, the received signal models of both linear memoryless MIMO channels and dispersive MIMO channels exhibiting memory, as well as the complex-valued versus real-valued MIMO system models. Then, an extensive review of the representative MIMO detection methods conceived during the past fifty years (1965-2015) is presented, and relevant insights as well as lessons are inferred for the sake of designing complexity-scalable MIMO detection algorithms that are potentially applicable to LS-MIMO systems. Furthermore, we divide the LS-MIMO systems into two types, and elaborate on the distinct detection strategies suitable for each of them. The type-I LS-MIMO corresponds to the case where the number of active users is much smaller than the number of BS antennas, which is currently the mainstream definition of LS-MIMO. The type-II LS-MIMO corresponds to the case where the number of active users is comparable to the number of BS antennas. Finally, we discuss the applicability of existing MIMO detection algorithms in LS-MIMO systems, and review some of the recent advances in LS-MIMO detection.

Index Terms—Co-channel interference (CCI), equalization, large-scale/massive MIMO, multiuser detection, MIMO detection.

Glossary

3G third generation.
4G fourth generation.
5G fifth generation.
A-CPDA approximate complex-valued probabilistic data association.
ACO ant colony optimization.
AME asymptotic-multiuser-efficiency.
APP a posteriori probability.
ASIC application-specific integrated circuit.
AWGN additive white Gaussian noise.
BALM block alternating likelihood maximization.
BC-SDPR bound-constrained semidefinite programming relaxation.
BER bit-error rate.
BI-GDFE block-iterative generalized decision feedback equalizer.
BLER block-error rate.
BP belief propagation.
BPSK binary phase-shift keying.
BS base station.
CAGR compound annual growth rate.
CCI co-channel interference.
CDMA code-division multiple-access.
CLPS closest lattice-point search.
CMOS complementary metal-oxide semiconductor.
CPDA complex-valued probabilistic data association.
CR cognitive radio.
DFD decision-feedback detector.
DMT diversity-multiplexing tradeoff.
DS-CDMA direct-sequence code-division multiple-access.
DSNR decreasing signal-to-noise ratio.
EB exabytes.
EM expectation-maximization.
EXIT extrinsic information transfer.
FCSD fixed-complexity sphere decoding/decoder.
FDM frequency-division multiplexing.
FDMA frequency-division multiple-access.
FEC forward-error-correction.
FER frame-error rate.
FH-CDMA frequency-hopped code-division multiple-access.
FIR finite impulse response.
GA genetic algorithm.
GSNR greatest signal-to-noise ratio.
HNN Hopfield neural network.
LAI interantenna interference.
IC integrated circuit.
ICI interchannel interference.
IDD iterative detection and decoding.
IMSE increasing mean-square error.
ISI intersymbol interference.
JPDA joint probabilistic data association.
LAS likelihood ascent search.
LDPC low-density parity-check.
LLL Lenstra-Lenstra-Lovász.
LMSE least mean-square error.
LR lattice-reduction.
LS least-squares.
LS-MIMO large-scale multiple-input multiple-output.
LSD list sphere decoding.
LTE long-term evolution.
LTE-A Long Term Evolution-Advanced.
M2M machine-to-machine.
MAI multiple-access interference.
MAME maximum asymptotic-multiuser-efficiency.
MAP maximum a posteriori.
MBER minimum bit error rate.
MC-CDMA multicarrier code-division multiple-access.
MED minimum Euclidean distance.
MF matched filter.

The financial support of the Research Councils UK (RCUK) under the India-UK Advanced Technology Center (IU-ATC), of the EU under the auspices of the Concerto project, and of the European Research Councils Advanced Fellow Grant is gratefully acknowledged.

The authors are with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK (e-mail: {lh, shaoshi.yang}@ecs.soton.ac.uk).
The multimedia data traffic conveyed by the global mobile networks has been soaring [1]–[5], and this trend is set to continue, as indicated by Cisco’s visual networking index (VNI) forecast [6], [7]. More specifically, the global mobile data traffic grew 81% in 2013, up from 0.82 exabytes (EB), i.e. $0.82 \times 10^{18}$ bytes per month at the end of 2012 to 1.5 EB per month at the end of 2013; furthermore, as predicted in Fig. 1, it will increase nearly 11-fold between 2013 and 2018, which translates to a compound annual growth rate (CAGR) of 61% for the period spanning from 2013 to 2018, reaching 15.9 EB per month by 2018 [6], [7]. As seen from Fig. 2, this explosive growth is mainly fuelled by the prevalence of smartphones, laptops and tablets, as well as by the emergence of machine-to-machine (M2M) communications [8]–[19]. Additionally, the design of wireless communication systems is highly constrained by the paucity of radio spectrum, which is exemplified by the overcrowded frequency allocation chart of the United States [20]. As a consequence of the combined effect of the mobile data traffic growth trend and the scarcity of favorable radio spectrum in the low-loss frequency-range, the forthcoming fifth generation (5G) communication systems have to resort to the employment of massive/large-scale multiple-input multiple-output (LS-MIMO) transmission techniques, which invoke a large number of antenna elements at the transmitter and/or receiver for achieving a high spectral-efficiency [21]–[29] and high energy-efficiency [25], [30]–[32].

A range of other fundamental technologies conceived for 5G communications are closely related to LS-MIMO. For example, both millimetre wave communications [33] and LS-MIMOs may be regarded as enabling techniques facilitating high-dimensional physical-layer communication technologies. Their difference is that LS-MIMOs achieve high dimensionality in the frequency domain by operating at frequencies ranging from about 30 GHz to 300 GHz, which is much higher than the operating frequencies of contemporary third generation (3G)/4G systems.
Fig. 2. Cisco VNI: global mobile devices and connections growth forecast, 2013-2018.

Fig. 3. There is a natural marriage amongst LS-MIMO, millimetre wave, and small-cell based HetNet, which constitute three fundamental technologies for 5G wireless communications.

(from 450 MHz to 3.5 GHz). Furthermore, owing to the much shorter wavelength, millimetre wave technologies may facilitate compacting a large number of antenna elements in a relatively small space. Additionally, the coverage area of a single cell of millimetre wave communication systems may be significantly smaller than a single cell of 3G/4G systems. As a result, small-cell based heterogeneous network (HetNet) architecture is required. Therefore, as shown in Fig. 3, there is a natural marriage amongst LS-MIMO, millimetre wave and small-cell technologies.

B. Why is MIMO Detection Important and Challenging?

As Claude Shannon pointed out, “The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point” [34]. Compared to conventional single-input single-output systems, e.g. the single-antenna point-to-point system, in MIMO systems we have multiple interfering messages/symbols transmitted concurrently, and at the receiver these symbols are expected to be detected/decoded subject to the contamination of random noise or interference, as shown in Fig. 4. The multiple symbols may be detected separately or jointly. As opposed to separate detection, in joint detection each symbol has to be detected taking into account the characteristics of the other symbols. As a beneficial result, typically joint detection is capable of achieving a significantly better performance than separate detection, although joint detection imposes a higher computational complexity.

The joint detection of multiple symbols in MIMO systems is of central importance for the sake of realizing the substantial benefits of various MIMO techniques. This is because the co-channel interference (CCI) routinely encountered in MIMO-based communication systems constitutes a fundamental limiting characteristic [24], [35]–[44]. Unfortunately, the optimum MIMO detection problem was proven non-deterministic polynomial-time hard (NP-hard) [45]–[47], thus all known algorithms conceived for solving the problem optimally have a complexity exponentially increasing with the number of decision variables. As a result, the computational complexity of the optimum maximum-likelihood (ML) criterion or the maximum a posteriori (MAP) criterion based MIMO detection algorithms quickly become excessive as the number of decision variables increases. Thanks to the rapid develop-
ment of the semiconductor industry, the hardware computing power has been dramatically increasing over the years, and in some cases a “not-so-extreme” computational complexity is no longer regarded as a bottleneck of practical applications. However, it should be noted that while transistors get faster and smaller, supply voltages cannot be reduced significantly in modern complementary metal-oxide semiconductor (CMOS) processes. Therefore, virtually all modern integrated circuits (ICs) encounter an integration density limit owing to the maximum tolerable internal temperature imposed by the excessive power consumption or power density. In other words, this power bottleneck still limits today’s IC development. As a consequence, one cannot simply rely on Moore’s law, and even modest-complexity MIMO detection algorithms could be too power hungry for battery-powered devices. Hence low-complexity, yet high-performance suboptimum MIMO detection algorithms are needed for practical MIMO applications.

C. The Contributions of This Paper

In this paper, an extensive review of the family of representative MIMO detection methods invented during the past fifty years is presented in a unified mathematical model\(^2\), although practical MIMO schemes have various subtleties. Our particular focus is on complexity-scalable MIMO detection algorithms potentially applicable to LS-MIMO systems [25]. The algorithms surveyed include the classic linear MIMO detection, the interference-cancellation based MIMO detection, the tree-search based MIMO detection, the lattice-reduction (LR) aided MIMO detection, the probabilistic data association (PDA) based MIMO detection, and the semidefinite programming relaxation (SDPR) based MIMO detection. Several high-quality books or reviews were published on MIMO detection [48]–[52]. They were predominantly dedicated to CDMA systems in the 1990s [48]–[50] or to conventional small-/medium-scale MIMO systems [51], [52], whereas LS-MIMOs just became a hot research topic at the time of writing [25]. On the other hand, they mainly covered the most common suboptimum MIMO detection methods, such as the linear zero-forcing (ZF) detector, the linear minimum mean-square error (MMSE) detector and various interference-cancellation based detectors [48]–[51], or focused largely on a single type of MIMO detector [52]. In comparison, there is a paucity of reviews on more advanced MIMO detection methods, such as the tree-search based MIMO detectors (the sphere decoder (SD) constitutes an instance of the tree-search based MIMO detectors) [53]–[82], the LR based MIMO detectors [83]–[103], as well as the PDA [104]–[133] and the SDPR [134]–[147] based detectors etc., although a concise tutorial on some of these detectors was given in [148]. Additionally, most of the existing research on LS-MIMO is focused on the precoding/beamforming based downlink of a special case of LS-MIMO, where one side of the communication link has a significantly higher number of antennas than the other. By contrast, only limited attention has been dedicated to the uplink of general LS-MIMOs. Hence, our goal is to fill these gaps in the open literature. For the sake of clarity, the organization of this paper is shown in Fig. 5.

II. The Nature of Co-Channel Interference

To gain profound insights into the intricacies of the MIMO detection problem, let us briefly reflect on the nature of the CCI in this section. The nature of CCI depends on the specific context. In this paper, it is defined in its most generic form as the interfering signal imposed by multiple transmissions taking place on channels which are mutually non-orthogonal. Mathematically, CCI may also be interpreted as interfering signals that span a subspace having a “non-empty” intersection with the subspace spanned by the desired signals. The channel-induced non-orthogonality may be observed in the frequency, time and/or space domain, as shown in Fig. 6. To recover the desired signal at the receiver, the desired signal has to be distinguishable from the interference in at least one domain. In the extreme case, if the multiple transmissions are highly non-orthogonal in all domains, then it may become impossible to recover the desired signal by any means.

In essence, the CCI originates from signal-feature-overlapping of multiple transmissions. For example, in spectrum-efficient communication systems such as the code-division multiplexing /multiple-access (CDM/CDMA) systems [149]–[151] and the space-division multiplexing /multiple-access (SDM/SDMA) systems [21], [44], [152]–[158], multiple transmissions are often deliberately arranged to take place simultaneously over the same frequency band. These “frequency sharing” and “time sharing” strategies result in a “frequency-overlapping” and a “time-overlapping” phenomenon, respectively. It is worth pointing out that as far as radio waves are concerned, rigorously the CCI always tends to exist in the frequency, time and space domains. For example, when no deliberate frequency-overlapping is arranged, the “frequency-overlapping” is due to the underlying fact that for all realizable, time-limited radio waves, their absolute bandwidth is infinite [159], [160], as shown in Fig. 7. In other words, every active radio transmitter has an impact on every operating radio receiver. Similarly, for a strictly bandwidth-limited signal, its time duration has to be infinite. With respect to the space domain, it is well known that the propagation of electromagnetic energy in free space is determined by the inverse square law [21], [158], [161], i.e.

\[
S = \frac{P_i}{4\pi d^2},
\]

where \( S \) is the power per unit area or power spatial density (in Watts per metre-squared) at distance \( d \), and \( P_i \) is the total power transmitted (in Watts). Hence, theoretically, the radio signals cannot be stopped, they are only attenuated in the frequency, time and space domains.

In engineering practice, fortunately, by using well-designed filters [162], [163], typically the waveform of the time-limited signal can be shaped so that most energy of the signal can be kept within a given limited frequency-band, and thus the signal energy leakage outside the target frequency-band can be reduced to a sufficiently low level. Similarly, in the space

\(^2\)This means that the algorithms conceived for the equalization, multiuser detection and multi-antenna detection problems can be treated under the same umbrella of the MIMO detection model of (1). More discussions on the similarities and differences amongst these three problems are provided in Section III and Section IV, as well as are found in the first two paragraphs of Section VIII and the last but one paragraph of Section VIII-B.
in these three domains is a result of a deliberate design. In this context, the signals are made as much distinguishable as possible in one domain, and as much overlapping as possible in the remaining domains. Our task is to recover the desired signal based on this deliberate arrangement.

Since the frequency, time and space domains represent the fundamental physical features of signal transmission, each of them corresponds to a distinct multiplexing/multiple-access scheme [164], namely the frequency-division multiplexing/multiple-access (FDM/FDMA), time-division multiplexing/multiple-access (TDM/TDMA), and SDM/SDMA, respectively. All of these multiple-access techniques aim for avoiding CCI by orthogonalizing the channel access in a certain domain. It is worth noting that compared to these three fundamental domains, the spreading code sequences used in CDM/CDMA systems do not constitute an independent domain. This is because the orthogonality of the spreading code sequences is essentially a special case of time-domain orthogonality. In systems using spreading codes, in principle we pursue to transmit orthogonal code sequences to minimize the inter-code interference. Although it is mathematically possible to construct perfectly orthogonal code sequences, the orthogonality of these code sequences is typically degraded in practical transmissions [151]. Moreover, since the number of theoretically orthogonal code sequences is rather limited, often quasi-orthogonal code sequences are adopted in practice [151]. Therefore, typically substantial interference is imposed by the non-orthogonality
propagation channel, hence typically substantial interference
physical size of transmitters/receivers and by the random
resolution of space-division tends to be undermined by the
may be relatively easy to obtain [164]. By contrast, the
"spreading codes" in frequency- and time-division systems
practice, by using guard intervals in the corresponding domain,
in these three domains. More specifically, in
access resolution
degree of differences in terms of
domain, respectively. Note, however, that there exist a certain
degree of differences in terms of their multiplexing/multiple-
access resolution in these three domains. More specifically,
in practice, by using guard intervals in the corresponding domain,
a good resolution of frequency-division and time-division may
be readily maintained – in other words, the orthogonality of
“spreading codes” in frequency- and time-division systems
may be relatively easy to obtain [164]. By contrast, the
resolution of space-division tends to be undermined by the
physical size of transmitters/receivers and by the random
propagation channel, hence typically substantial interference
is imposed by the non-orthogonality of “spreading codes” in
practical space-division systems [21], [44], [152]–[158]. This
is similar to the case in CDM/CDMA systems and explains
why MIMO detection typically represents a more significant
problem in CDM/CDMA systems and SDM/SDMA systems
than in FDM/FDMA and TDM/TDMA systems.

Additionally, there are more advanced multicarrier based
orthogonal multiple-access techniques, such as orthogonal
frequency-division multiple-access (OFDMA) [166]–
[168], single-carrier frequency-division multiple-access (SC-
FDMA) [169], and multicarrier CDMA [170]–[177]. Despite
their potential advantages in averaging interference over dif-
different subcarriers for different users, they usually suffer high
sensitivity to frequency offset, which leads to intercarrier in-
terference. Therefore, judicious frequency offset compensation
scheme and frequency reuse scheme have to be designed to
minimize the intercarrier interference.

In this paper, the CCI considered mainly refers to the
interference in SDM/SDMA or CDM/CDMA systems, where
multiple transmissions often take place simultaneously, or
partially simultaneously over the same frequency. Depending
on specific applications, CCI is often alternatively termed as
intersymbol interference (ISI), interchannel interference (ICI),
interantenna interference (IAI), multiuser interference (MUI),
multiple-access interference (MAI), and multiple-stream inter-
fERENCE (MSI) etc.

III. CONCEPT AND GENERALITY OF MIMO DETECTION

As a family of general techniques for physical-layer CCI
management, MIMO detection deals with the joint detection
of several information-bearing symbols transmitted over a
communication channel having multiple inputs and multiple
outputs. This problem is of fundamental importance for mod-
ern high-throughput digital communications. Rigorously, the
MIMO detection problem arises if and only if the respective
subchannels of the multiple inputs are not orthogonal to
each other, and hence there exists interference between the
outputs. As a generic mathematical model, the MIMO detec-
tion problem underpins numerous relevant applications, while
the physical meaning of the inputs and outputs herein may
vary in different contexts. For instance, in band-limited ISI
channels requiring equalization, the inputs refer to a sequence
of symbols interfering with each other, and the outputs are
the signals received within a given observation window in
the time/frequency domain [178]. In single-user SDM-MIMO
systems equipped with multiple transmit and receive antennas
[179]–[181], the inputs refer to the vector of modulated
symbols that are transmitted from multiple collocated transmit
antennas, while the outputs refer to the vector of received
signals recorded at multiple collocated receive antennas, as
shown in Fig. 8. This is indeed a canonical scenario of inves-
tigating MIMO detection algorithms [148]. A third example
is the uplink of multiuser multiple-antenna systems [182],
[183], where the inputs may be multiple transmitted symbols
belonging to a cluster of geographically distributed single-

Note that the so-called SC-FDMA is actually a multicarrier multiple-access
technique, although somewhat misleadingly it has “single-carrier” in its name.
antenna mobile stations (MSs), and the outputs may be the signals received at the serving base station (BS) equipped with multiple colocated antennas, as shown in Fig. 9. This is actually the so-called SDMA system [21], [44], [152]–[158]. Yet another important example represented by Fig. 9 is the uplink of CDMA systems [48]–[51], where the inputs are the transmitted symbols of distributed single-antenna MSs, and the outputs are typically generated by filtering the signal received at the single-antenna BS with a bank of matched filters (MFs), whose impulse responses are matched to a set of a priori known user-signature waveforms.

Here, it should be emphasized that whether the multiple inputs and/or the multiple outputs are “collocated or not” is extremely important in determining the signal processing techniques to be used. If multiple inputs/outputs are colocated, the cooperative joint encoding/decoding of the inputs/outputs can be conducted [21], [44], [158], [184]–[190], which renders joint MIMO transmission/detection feasible. For example, the single-user MIMO system shown in Fig. 8 has both its transmit and receive antennas collocated, hence it enjoys the privilege of performing both joint encoding and joint decoding. As a benefit, both simultaneous transmission and simultaneous reception can be attained relatively simply. By contrast, the multiple-access MIMO system of Fig. 9 is typically not capable of joint encoding at the user side, hence the uplink transmissions of both CDMA and SDMA systems are asynchronous by nature.

Additionally, as far as the downlink of multiuser MIMO systems, namely the multiuser MIMO broadcast channel of Fig. 10 is concerned, typically most of the sophisticated signal processing tasks are conducted in the form of transmit preprocessing (i.e. precoding) at the BS, where colocated inputs are available for cooperative joint encoding [191]–[196]. As a result, detection at the user becomes less challenging. Since the investigation of MIMO transmit preprocessing techniques is beyond the scope of this paper, we will not elaborate on it in the sequel.

Finally, when both the transmitters and the receivers are geographically distributed, the MIMO channel turns into either an interference channel [197]–[205] or an X channel [206]–[209], which are shown in Fig. 11 and Fig. 12, respectively. An interference channel characterizes a situation where each transmitter, potentially equipped with multiple antennas, only wants to communicate with its dedicated receiver, and each receiver, possibly equipped with multiple antennas as well, only cares about the information arriving from the corresponding transmitter. There is a strict one-to-one correspondence between the multiple transmitters and the multiple receivers. Therefore, each transmission link interferes with the others. By comparison, in the MIMO X channel relying on L transmitters and K receivers, each transmitter has an independent message for each receiver. Hence, there are a total of KL independent messages to deliver. The MIMO X channel is a more generalized model, which encompasses the MIMO multiple access channel of Fig. 9, the MIMO broadcast channel of Fig. 10 and the MIMO interference channel of Fig. 11 as its special cases. Despite their difference, the MIMO interference channel and X channel share a key common feature, namely they both have distributed transmitters and receivers. The distributed nature of transmitters and receivers makes the signal processing required for mitigating the detrimental effects of the MIMO interference channel and X channel far more challenging compared to the single-user MIMO channel. In fact, the capacity analysis and the signal processing techniques for MIMO interference channel and X channel still constitute a largely open field, and most of existing efforts have aimed for transforming the MIMO interference channel and X channel so that cooperation at the transmitter/receiver side can be exploited to some degree, at least in some specific scenarios. For example, in multicell systems, BS cooperation [131], [194], [210]–[219], also known as joint multicell processing [24], [194], has been advocated for the sake of transforming the MIMO interference channel and X channel to a number of cooperative multiuser MIMO channels. Additionally, the recent advances in the capacity analysis of the MIMO interference channel and X channel have stimulated significant interests in interference channel.
alignment [207], [208], [209], [220]–[225], which is essentially constituted by a family of precoding/beamforming techniques for the MIMO interference channel and X channel. The problems related to interference alignment are also beyond the scope of this paper and will not be discussed in detail.

IV. FORMAL DEFINITION OF THE MIMO DETECTION PROBLEM

Despite the fact that similar problems have been known for a while [46], [50], [51], [226]–[261], the term “MIMO detection” became widespread mainly with the advent of multiple-antenna techniques during the mid-1990s [152]–[156], [179]–[181], [262]–[273]. As a result, in the narrow sense, MIMO detection usually refers to the symbol detection problem encountered in narrow-band SDM based multiple-antenna systems, such as the vertical Bell Laboratories layered space-time (VBLAST) system [179]–[181]. However, we emphasize that as a family of important signal processing techniques, MIMO detection should be interpreted based on a generic mathematical model, as detailed below.

In the generic sense, the MIMO detection problem can be defined for an \( N_I \)-input linear system whose transfer function is described by a matrix having non-orthogonal columns and its \( N_O \) outputs are contaminated by additive random noise. Note that the noise does not necessarily obey the Gaussian distribution. The multiple inputs can be denoted as a vector \( s \), which is randomly drawn from the set \( A^{N_I} \) composed by \( N_I \)-element vectors, whose components are from a finite set \( A \) and the \textit{a priori} probability of selecting each vector from \( A^{N_I} \) is identical. The set \( A \) is usually referred to as the constellation alphabet, whose elements can take either real or complex values. Additionally, \( \hat{s}_n \), \( n = 1, \ldots, M^{N_I} \), represents the realizations of \( s \), hence they are the elements of \( A^{N_I} \). Then the relationship between the inputs and the outputs of this linear system can be characterized by

\[
y = Hs + n,
\]

where \( y \in \mathbb{R}^{N_O} \) is the received signal vector, \( H \in \mathbb{R}^{N_O \times N_I} \) is the transfer function/channel matrix of the system, and \( n \in \mathbb{R}^{N_O} \) represents the additive random noise vector. Depending on the specific applications considered, \( F \) can be either the field of real numbers, \( \mathbb{R} \), or the field of complex numbers, \( \mathbb{C} \). Concisely speaking, any system having multiple inputs, multiple outputs and subject to additive random noise can be regarded as a MIMO system, but the MIMO detection problem considered herein is only confronted in MIMO systems whose channel matrix is non-orthogonal in columns. It is worth noting that the constellation alphabet \( A \), the number of inputs \( N_I \) and the number of outputs \( N_O \) are typically regarded as constant quantities\(^6\) for a given system. Hence, they are assumed to be known by default, although this will not be explicitly emphasized, unless necessary. As a further note, when the input symbol vectors of multiple consecutive time slots are associated with each other via space-time coding [266]–[274], the MIMO system model is given by

\[
Y = HC + N,
\]

where \( Y \) is a matrix denoting the signal received in multiple time slots, \( C \) is a matrix representing the space-time codeword, and \( N \) is the corresponding signal noise matrix. We can obtain (1) from (2) by setting the number of time slots considered to one. In this regard, (2) is more general than (1). However, (2) is mainly used for characterizing space-time coding aided MIMO systems, where typically the MIMO detection problem defined in this paper does not exist. This is because the optimal ML decoding can be simply implemented using the separate symbol-by-symbol decoding strategy [for orthogonal space-time block codes (STBCs)] or the pairwise decoding strategy (for quasi-orthogonal STBCs) [274]. Therefore, in most cases associated with MIMO detection, we rely on the system model (1).

Based on the generic mathematical model of (1), the basic task of MIMO detection is to estimate the input vector \( s \) relying on the knowledge of the received signal vector \( y \) and the channel matrix \( H \). Note that for \( y \), typically its exact value has to be known, while for \( H \), sometimes only the knowledge of its statistical parameters is available. To elaborate a little further, if the \textit{instantaneous} value of \( H \) is known from \textit{explicit} channel estimation, the detection of \( s \) is said to rely on \textit{coherent} detection. By contrast, if the explicit estimation of the instantaneous channel state is avoided, the detection of \( s \) belongs to the family of \textit{noncoherent} detection schemes. In the latter case, the channel estimation is either performed implicitly in signal detection, or it is completely

---

\(^4\)Note that the first multi-antenna based MIMO system was attributed to a patent granted to Paulraj and Kailath in 1994 [262]. Gerlach [153], Roy [152], [155] and Ottersten [152], [154] initiated the earliest research on SDMA systems. The earliest contribution demonstrating the huge capacity of multi-antenna based MIMO systems may be attributed to Telatar [263], [264], as well as Foschini [179], [265] and Gans [265], followed by other members of the team at Bell Labs [180], [181]. On the other hand, Tarokh, Jafarkhani, Calderbank, Naguib et al. [266]–[272] as well as Alamouti [273] are pioneers of space-time code design.

\(^5\)In the additive random noise contaminated ISI channels, the received signal is given by \( y_n = \sum_{i=0}^{M} h_i s_{n-i} + w_n \), \( n = 0, 1, \ldots, L - 1 \), or by \( y = Hs + w \), where \( y = [y_0, y_1, \ldots, y_{L-1}]^T \), \( w = [w_0, w_1, \ldots, w_{L-1}]^T \), \( L \) is the number of symbols observed, and \( M_h \) is the memory length of the ISI channel.

\(^6\)However, in an adaptive system both the constellation alphabet \( A \) and the number of inputs \( N_I \) might be varying. But this adaptation is typically constrained by a discrete size-limited codebook.
avoided, whereas typically the statistical knowledge of the channel matrix $H$ is invoked for supporting signal detection. Additionally, the noncoherent MIMO detection schemes usually require that the input symbols are subject to some form of differential encoding, which imposes correlation on the input symbols, and as a result, typically a block-by-block based sequence detection has to be employed. This is the so-called multiple-symbol differential detection [274]–[283], which usually leads to higher computational complexity than the symbol-by-symbol based detectors of coherent MIMO systems. Moreover, the noncoherent detectors typically exhibit degraded power efficiency, which results in an inherent performance loss compared to their coherent counterparts, unless the block size is sufficiently large. Therefore, we have to consider the performance-versus-complexity tradeoff in choosing the proper block size. However, similar to the coherent detection of STBCs, there exist simple symbol-by-symbol or pairwise noncoherent detection schemes [276], [277] for differential space-time modulation. As a result, the decoding complexity of the differential space-time modulation increases linearly, instead of exponentially, with the number of antennas [274]. In this paper, we focus our attention on coherent MIMO detection. Then, from the perspective of mathematical mapping, a coherent MIMO detector is defined as:

$$\hat{s} = \mathcal{D}(y, H) : \mathbb{R}^{NO} \times \mathbb{R}^{NO \times N_1} \mapsto \mathbb{K}^{N_1},$$

where $\hat{s}$ is the estimate of $s$.

V. MIMO SYSTEM MODEL FOR LINEAR MEMORYLESS CHANNELS

Bearing in mind specific applications, the system model of (1) may be established either in the time domain or in the frequency domain, and may be applied to both memoryless channels and dispersive channels exhibiting memory [21], [44], [50], [51], [158], [261]. With respect to linear memoryless MIMO channel, a canonical example is the narrowband single-carrier synchronous VBLAST-style SDM-MIMO system [179]–[181] communicating over flat fading channels, as shown in Fig. 13. Because the system’s outputs at the current time interval are independent of the system’s inputs at previous time intervals, its baseband equivalent discrete-time (i.e. sampled) system model, representing an instance of the generic model (1), can be written as

$$\begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_{N_r}
\end{bmatrix} =
\begin{bmatrix}
  h_{1,1} & h_{1,2} & \cdots & h_{1,N_t} \\
  h_{2,1} & h_{2,2} & \cdots & h_{2,N_t} \\
  \vdots & \vdots & \ddots & \vdots \\
  h_{N_r,1} & h_{N_r,2} & \cdots & h_{N_r,N_t}
\end{bmatrix}
\begin{bmatrix}
  s_1 \\
  s_2 \\
  \vdots \\
  s_{N_t}
\end{bmatrix}
+ \begin{bmatrix}
  n_1 \\
  n_2 \\
  \vdots \\
  n_{N_r}
\end{bmatrix}.$$  

In this specific application, we have $N_t = N_I$ and $N_r = N_O$, which represent the number of transmit and receive antennas, respectively. Furthermore, $h_{j,i}$ denotes the (complex-valued) impulse response between the $i$th transmit antenna and the $j$th receive antenna, with $i = 1, 2, \cdots, N_t$ and $j = 1, 2, \cdots, N_r$. Another example is the multiple-antenna aided orthogonal frequency-division multiplexing (OFDM) system [283] communicating over frequency-selective channels, where each subcarrier subjected to a specific frequency-domain attenuation is narrowband and the model (4) applies to each subcarrier. Notably, for this linear memoryless MIMO channel, the one-shot detection which relies only on a single received signal vector $y = [y_1, y_2, \cdots, y_{N_r}]^T$ is adequate. Additionally, for the sake of fair comparison with the single-input single-output systems, typically the energy normalization of $\mathcal{E}(s_i) = 1$ or $\mathcal{E}(s) = 1$ is imposed on the transmitted symbols.

VI. MIMO SYSTEM MODEL FOR DISPERSIVE CHANNELS EXHIBITING MEMORY

On the other hand, when considering the stand-alone wideband VBLAST system\(^\dagger\) communicating over frequency-selective MIMO channels [112], [285]–[288], the link between each input-output pair may be modelled by a linear finite impulse response (FIR) dispersive channel, whose sampled version can be denoted as the (possibly complex-valued) vector $h_{j,i} = (h_{j,i,1}, h_{j,i,2}, \cdots, h_{j,i,L})^T$. Here $L$ is the maximum number of multipath components in each link, and it is also known as the channel memory length. In this case, the one-shot detection which utilizes a single $N_r$-element received signal vector is not optimal. Instead, the sequence detection using multiple $N_r$-element received vectors has to be used.

We assume that a block-based transmission structure relying on zero-padding for eliminating the interblock interference is used, which is beneficial for alleviating the performance degradation imposed by noise enhancement or error propagation [289]. Following zero-padding, a transmission block becomes a frame which occupies $K = N + P$ sampling intervals, where $N$ is the number of sampling intervals occupied by information-bearing symbol vectors in the frame, while $P \geq L - 1$ represents the number of sampling intervals during which $P$ consecutive $N_r$-element zero vectors are

\(^\dagger\)Multicarrier techniques, such as OFDM, are not used in this system. However, when MIMO-OFDM systems [284] are considered, the MIMO detection is carried out on each subcarrier separately.
inserted at the tail of the frame. Here we set \( P = L - 1 \). Given the above-mentioned transmitted frame, the entire received signal vector may be generated by a concatenation of \( K \) noise-contaminated sampled received signal vectors, namely, \( \mathbf{y} = (\mathbf{y}^T[0], \mathbf{y}^T[1], \ldots, \mathbf{y}^T[K - 1])^T \), where \( \mathbf{y}[k] = (y_1[k], y_2[k], \ldots, y_{N_r}[k])^T \) represents the \( N_r \) outputs at the \( k \)th sampling instant, \( k = 0, 1, \ldots, K - 1 \). Then, the baseband signal received by the \( j \)th receive antenna at the \( k \)th sampling instant is given by

\[
y_j[k] = \sum_{i=1}^{N_t} \sum_{l=0}^{L-1} h_{j,i}^l s_i[k - l] + n_j[k],
\]

where \( h_{j,i}^l \), the \( l \)th element of \( \mathbf{h}_{j,i} \), \( l = 0, 1, \ldots, L - 1 \), denotes the channel gain of the \( l \)th path between the \( i \)th transmit antenna and the \( j \)th receive antenna. Furthermore, \( s_i[k] \) is the symbol transmitted from the \( i \)th transmit antenna at the \( k \)th sampling instant, and \( n_j[k] \) represents the noise imposed on the \( j \)th receive antenna at the \( k \)th sampling instant. Similar to \( \mathbf{y}[k] \), we define \( \mathbf{s}[k] = (s_1[k], s_2[k], \ldots, s_{N_t}[k])^T \) and \( \mathbf{n}[k] = (n_1[k], n_2[k], \ldots, n_{N_r}[k])^T \). Additionally, similar to \( \mathbf{y} \), we may construct \( \mathbf{s} = (s^T[0], s^T[1], \ldots, s^T[N - 1])^T \) and \( \mathbf{n} = (n^T[0], n^T[1], \ldots, n^T[K - 1])^T \). Then, the received signal corresponding to a transmitted frame can also be written following the matrix notation of (1), where the size of \( \mathbf{y} \) and \( \mathbf{n} \) is \( N_O = K N_r = (N + L - 1) N_r \), while that of \( \mathbf{s} \) is \( N_I = N N_t \), and the MIMO channel matrix \( \mathbf{H} \) has the banded Toeplitz structure [289] of:

\[
\mathbf{H} = \begin{bmatrix}
H^0 & 0 & \cdots & 0 \\
\vdots & H^0 & \ddots & \vdots \\
H^{L-1} & \ddots & \cdots & 0 \\
0 & H^{L-1} & \cdots & H^0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & H^{L-1}
\end{bmatrix},
\]

whose dimension is \((K N_r \times N N_t)\), and each entry \( \mathbf{H}^l \) of (6) is an \((N_r \times N_t)\)-element matrix containing the channel gains between all pairs of transmit and receive antennas for the \( l \)th path, i.e. we have

\[
\mathbf{H}^l = \begin{bmatrix}
h_{1,1}^l & h_{1,2}^l & \cdots & h_{1,N_t}^l \\
h_{2,1}^l & h_{2,2}^l & \cdots & h_{2,N_t}^l \\
\vdots & \vdots & \ddots & \vdots \\
h_{N_r,1}^l & h_{N_r,2}^l & \cdots & h_{N_r,N_t}^l
\end{bmatrix}.
\]

It is worth noting that the linear memoryless MIMO model and the linear MIMO model exhibiting memory may also be used for characterizing the family of synchronous and asynchronous CDMA systems, respectively. Additionally, the asynchronous CDMA systems can also be characterized by (1) in the \( z \) domain [50].

VII. COMPLEX-VALUED VERSUS REAL-VALUED MIMO SYSTEM MODEL

As we mentioned in Section IV, the generic MIMO system model of (1) can be defined both in the field of real numbers, \( \mathbb{R} \), and in the field of complex numbers, \( \mathbb{C} \). Since the complex-valued modulation constellations, such as quadrature amplitude modulation (QAM) and phase-shift keying (PSK), are often employed in digital communications, the complex-valued MIMO system model is typically a natural and more concise choice for the formulation and performance analysis of the algorithms considered.

The complex-valued and the real-valued MIMO system models are often mutually convertible. More specifically, if we assume that the generic MIMO system model of (1) is defined in \( \mathbb{C} \), and assume that the real part and the imaginary part of \( \mathbf{s} \) are uncorrelated,\(^8\) then the complex-valued MIMO system model of (1) can be transformed to an equivalent real-valued system model of

\[
\tilde{\mathbf{y}} = \tilde{\mathbf{H}} \tilde{\mathbf{s}} + \tilde{\mathbf{n}},
\]

where \( \tilde{\mathbf{y}} = \begin{bmatrix} \Re(\mathbf{y}) \\Im(\mathbf{y}) \end{bmatrix}, \tilde{\mathbf{s}} = \begin{bmatrix} \Re(\mathbf{s}) \\Im(\mathbf{s}) \end{bmatrix}, \tilde{\mathbf{n}} = \begin{bmatrix} \Re(\mathbf{n}) \\Im(\mathbf{n}) \end{bmatrix} \), and \( \tilde{\mathbf{H}} = \begin{bmatrix} \Re(\mathbf{H}) & -\Im(\mathbf{H}) \\
\Im(\mathbf{H}) & \Re(\mathbf{H}) \end{bmatrix} \).

However, the above real-valued decomposition is only applicable to MIMO systems employing real-valued constellation or rectangular QAM constellations (but not for PSK, star QAM [290]–[293] and near-Gaussian constellations [294], [295] etc.), which severely limits its applicability. Furthermore, in many applications, complex-valued operations are more efficient for hardware implementation. The reason for this fact is twofold. Firstly, decomposing an \((N_I \times N_I)\)-element MIMO system into a real-valued system requires storage of the \((2 N_I \times 2 N_I)\)-element real-valued channel matrix \( \tilde{\mathbf{H}} \), which is twice larger than having \( 2 N_I \) real-valued elements that would be needed for a standard representation of the original complex-valued channel matrix \( \mathbf{H} \). Secondly, implementing complex-valued arithmetics in hardware (e.g. very-large-scale integration (VLSI) based circuits) is straightforward and does not result in more complex hardware. For example, a complex-valued multiplier can be built using 4 real multipliers and 2 real adders, because we have \((a + j b)(c + j d) = ac - bd + j(ad + bc)\), or using 3 real multiplications and 5 real additions, because alternatively we have \((a + j b)(c + j d) = ac - bd + j[(a + b)(c + d) - ac - bd]\), which is known as the Gaussian technique of multiplying complex numbers [296]. As a result, the complex-valued model imposes a lower silicon complexity than that required by the real-valued decomposition based model. Therefore, in many cases the real-valued decomposition can be detrimental and hence MIMO detector designers are typically in favor of the complex-valued system model, owing to its flexibility concerning the choice of constellations and its efficiency in VLSI implementations.

\(^8\)For rectangular QAM, this uncorrelatedness assumption is almost always adopted for ease of decoding, even in single-input single-output systems. In channel-coded systems, where the channel codes may introduce correlation between the coded bits. However, in such systems, typically an interleaver is invoked after the encoder, which mitigates the correlation.
On the other hand, the real-valued MIMO system model may also enjoy some advantages, such as the increased freedom of manipulation in signal processing. To elaborate a little further, as far as the achievable performance is concerned, in most cases signal processing algorithms based on the complex-valued model of (1) and the real-valued MIMO system model of (8) deliver an equivalent performance. For example, [264] showed the equivalence between the complex-valued and the real-valued MIMO system models in the derivation of the optimal ML detector and the MIMO channel capacity. However, this equivalence does not always hold. For example, it was shown in [297] that the real-valued VBLAST detector outperforms its complex-valued counterpart, owing to its additional freedom in selecting the optimum detection ordering. Hence, a beneficial performance gain may be gleaned from transforming the complex-valued system model to the double-dimensional real-valued system model. This is also true for the tree-search based MIMO detectors, which will be introduced in Section VIII-D, when they invoke symbol detection ordering. More generally, the key insight inferred here is that for all MIMO detection algorithms whose performance is related to detection ordering, the real-valued system model based formulation is capable of providing a better performance than its complex-valued counterpart. This gain is achieved at the expense of extra redundancy in storage of the channel matrix $\mathbf{H}$, and if the symbol detection ordering technique is invoked, this redundancy cannot be avoided.

Additionally, it is worth noting that the real-valued formulation of the complex-valued MIMO system model is not unique. For example, a pairwise real-valued MIMO system model was used in [298], [299], which was shown to result in a reduced complexity compared to the conventional real-valued MIMO system model of (8). A more comprehensive investigation of the complex-valued versus real-valued MIMO detectors was presented in [300].

Finally, we emphasize that if the complex-valued random signals considered are improper or noncircular, the more advanced complex-valued signal processing techniques of [301]–[303] that rely on additional statistics and tools for fully characterizing the complex-valued random signals have to be used. More specifically, for a complex-valued random vector $\mathbf{x}$, in addition to the conventional covariance matrix of

$$
C_{xx} = \mathbb{E}[(\mathbf{x} - \mu_x)(\mathbf{x} - \mu_x)^H],
$$

(9)

another second-order statistic, namely the pseudo-covariance matrix defined as

$$
\tilde{C}_{xx} = \mathbb{E}[(\mathbf{x} - \mu_x)(\mathbf{x} - \mu_x)^T],
$$

(10)

has to be introduced for fully describing the complex-valued random vector. For a proper complex-valued random vector, the pseudo-covariance matrix vanishes, which is formulated as $\tilde{C}_{xx} = \mathbf{0}$. This results in the fact that for a proper complex-valued random scalar, the real and imaginary parts must have the same variance and be uncorrelated. Additionally, a circularly (symmetric) complex-valued random variable has a probability distribution that is invariant under rotation in the complex plane, namely the distribution of $\mathbf{x}$ must be the same as the distribution of $e^{j\theta} \mathbf{x}$, where we have $\theta \in [0, 2\pi)$. The conventional signal processing techniques often assume, usually implicitly, that the complex-valued random signals are proper or circularly symmetric\(^3\). However, these assumptions may not be justified in some applications, hence the complex-valued signal processing techniques may in certain circumstances achieve a better performance [120], [130]–[133]. For more details on the complex-valued signal processing, please refer to [301]–[303].

VIII. HISTORY AND STATE-OF-THE-ART OF MIMO DETECTION

The research of MIMO detection is a broad and vibrant area. Its embryonic concept dates back to the 1960s. The earliest contribution on MIMO detection was sparked off in 1967 [226], when Shnidman considered the equalization problem of a bandwidth-limited pulse modulation system. This system was modelled with the aid of $M$ waveforms, each of which is amplitude-scaled and simultaneously transmitted over a single physical channel, which has $M$ outputs corresponding to each signal waveform. In order to eliminate both the ISI between the pulse train and the interference between different waveforms (also known as crosstalk), Shnidman formulated a generalized Nyquist criterion and proposed an optimum linear receiver. This landmark contribution was essentially inspired by the classic Nyquist’s problem [304], which aims for the joint optimization of the transmitter and receiver for the sake of combating the ISI when communicating over a conventional single-input single-output channel. Since then, the MIMO detection problem has been studied in the context of diverse applications and under possibly different names. This half-century history can be roughly divided into four periods, as seen in Fig. 14, namely the period of combating crosstalk in the context of the early single-user FDM/TDM systems (1960s – 1970s) [226]–[229], [232], the

\(^3\)For complex-valued Gaussian random vector $\mathbf{Z}$, circular symmetry implies that $\mathbf{Z}$ is zero mean and proper.
period of multiuser detection (MUD) during the prevalence of CDM/CDMA systems (1980s – 1990s) [46], [50], [51], [230], [231], [229]–[261], the period of joint symbol detection in the small-/medium-scale multiple-antenna systems (mid-1990s – mid-2000s) [53], [54], [57]–[64], [68]–[72], [80], [82], [104]–[126], [128], [129], [134]–[146], [179]–[181], [261], [305]–[324], and the period of symbol detection in the large-scale multiple-antenna systems [118], [128], [176], [325]–[341]. Diverse MIMO detectors have been proposed for meeting the requirements imposed by a multiplicity of applications. These MIMO detectors can be categorised from various perspectives, such as optimum/suboptimum, linear/nonlinear, sequential/one-shot, adaptive/non-adaptive, hard-decision/soft-decision, blind/non-blind, iterative/non-iterative, synchronous/asyncronous, coded/uncoded etc. Note that a detailed discourse on the application of soft-decision MIMO detectors in near-capacity turbo/iterative receivers was provided in [342], while coherent and noncoherent MIMO detectors in the context of the emerging “space-time shift keying (STSK)” based multicarrier MIMO systems was presented in [177]. The representative MIMO detectors considered in this paper are summarized in Fig. 15.

Owing to the similarities between the classic equalization problem encountered in channels imposing ISI and the generic MIMO detection problem defined by (1) and (3), it is not surprising that the techniques, which were found to be effective in combating ISI were also often extended to the context of MIMO detection problems [343]. Some of the equalization algorithms which have been adapted for MIMO detection include, but not limited to, the ML sequence estimation (Viterbi algorithm) [344]–[348] based equalization, linear ZF equalization [161], linear MMSE equalization [161], ZF/MMSE aided decision-feedback equalization [161], adaptive equalization [349], [350], blind equalization [351], [352] etc, as detailed below.

A. Optimum MIMO Detector

The earliest work on optimum MIMO detectors dates back to 1976, when van Etten [229] derived an ML sequence estimation based receiver for combating both ISI and interchannel interference (ICI) in multiple-channel transmission systems. Explicitly, he demonstrated that under certain conditions, the performance of the ML receiver asymptotically approaches the optimum. In 1983, this conventional wisdom was explicitly proven wrong by Verdú [235], [236] with the introduction of the optimal MUD in the context of asynchronous/synchronous Gaussian multiple-access channels shared by K users. The full analysis and derivation of the optimum MUD was reported later in [237], [354], demonstrating that there is, in general, a substantial gap between the performance of the conventional SUMF and the optimal MUD performance. Additionally, upon identifying the non-Gaussian nature of the MUI, Poor and Verdú [355] also designed nonlinear single-user detectors for CDMA systems operating in diverse scenarios such as weak interferers, high spreading gains and high signal-to-noise ratio (SNR). The performance of the ML based optimum MIMO detector has been analyzed in [237], [354], [356]–[360].

There does exist some situations where a bona fide application of the central limit theorem is feasible and hence the MUI can be rigorously proven to be asymptotically Gaussian. However, even if the MUI may be accurately modelled as a Gaussian variable, the SUMF is still not the optimal receiver. This is because the output of the MF for the desired user does not constitute a sufficient statistic in the presence of MUI [361]. In other words, the SUMF is optimal only in the context of the single-user channel contaminated by additive white Gaussian noise (AWGN). By contrast, in multiple-access systems, unless the multiplexed signals (after passing through the channel) are orthogonal, the outputs of the MFs corresponding to the interfering users contain valuable information which may be exploited for the detection of the symbol of interest, and hence more intelligent joint detection strategies capable of exploiting all MFs’ outputs are required for achieving better detection performance.

2) Optimum Decision Criteria: When designing optimal detectors/receivers for communication systems, it is usually necessary to clarify in what sense the word “optimum” is referred to. This is because the specific choice of an optimal detector/receiver is strongly dependent on the specific assumptions and criteria of “goodness”. An optimal detector/receiver is the one that best satisfies the given criterion of goodness under a given set of assumptions. If either the criterion or the assumptions change, typically the choice of the optimal detector/receiver also changes. If the assumptions used in the theoretical analysis are inconsistent with the conditions of the realistic environment considered, then it is possible that the theoretically optimal detector/receiver obtained fails to provide valid insights and results for the practically achievable performance and designs. Special attention has to be paid to the definition of “optimum”, since the theoretically optimal results obtained are mainly invoked as a benchmark or bound, against which any other results can be compared. There are many criteria of goodness. As far as the performance of the detectors/receivers used in communication systems

10In fact, van Etten’s pioneering companion papers on the optimum MIMO detector [229] and on the optimum linear MIMO detector [228] were included in the book The best of the best: Fifty years of communications and networking research [353], which was compiled by the IEEE Communications Society in 2007, and he is the only researcher who has two sole-author papers included in this selection.
is concerned, the *minimum error probability criterion* is of primary interest, and hypothesis testing as well as likelihood ratios are of great importance. In Bayesian inference, the optimum decision criterion which minimizes the error probability based only on the observed signals and a given set of hypotheses is the maximum *a posteriori* (MAP) criterion. The error probability in communication systems can be measured in multiple scales, such as bit-error rate (BER), symbol-error probability in communication systems can be measured using Bayes’ rule, the a posteriori probability, may be expressed as

\[ D_{\text{MAP}} : \hat{s} = \arg \max_{s \in A} p(y|s). \]  

(11)

Using Bayes’ rule, the *a posteriori* probability (APP) in (11) may be expressed as

\[ \Pr(s|y) = \frac{p(y|s) \Pr(s)}{p(y)} = \frac{p(y|s) \Pr(s)}{\sum_{s' \in A} p(y|s') \Pr(s')}, \]  

(12)

where \( \Pr(s) \) is the *a priori* probability of \( s \), and \( p(y|s) \) is the conditional probability density function (PDF) of the observed signal vector \( y \) given \( s \). The MAP criterion can be simplified when each vector in \( A \) has an identical *a priori* probability, i.e. we have \( \Pr(s) = 1/|A| \) for all realizations of \( s \), where \(|A|\) represents the number of elements, i.e. the cardinality of the constellation alphabet \( A \). Furthermore, considering the fact that \( p(y) \) is independent of which particular signal vector is transmitted, then the MAP detector of (11) becomes equivalent to the ML detector of

\[ D_{\text{ML}} : \hat{s} = \arg \max_{s \in A} p(y|s). \]  

(13)

Therefore, the MAP criterion is usually used in the iterative detection and decoding (IDD) aided receiver of forward-error-correction (FEC)-coded systems, where the *a priori* probabilities of the transmitted symbols, \( \Pr(s) \), may be obtained with the aid of a backward-and-forward oriented iterative information exchange between the signal detector and the channel decoder. By contrast, the ML criterion is usually used in FEC-uncoded systems, where the *a priori* probabilities of the transmitted symbols cannot be made available by the channel decoder. If \( n \) is AWGN, then we have

\[ p(y|s) \propto \exp(-\|y - Hs\|_2^2), \]  

(14)

where the symbol \( \propto \) represents the relationship “is proportional to”. Consequently, we have

\[ \max_{s \in A} p(y|s) \Leftrightarrow \min_{s \in A} \|y - Hs\|_2^2, \]  

(15)

where the symbol \( \Leftrightarrow \) represents the relationship “is equivalent to”. Therefore, the ML detection problem for the system model of (1) can be reformulated as the finite-set constrained least-squares (LS) optimization problem of

\[ \hat{s}_{\text{ML}} = \arg \min_{s \in A} \|y - Hs\|_2^2, \]  

(16)
which can also be interpreted as the minimum Euclidean distance (MED) criterion.

Note, however, that the above-mentioned MAP, ML and MED criterion based MIMO detectors all aim for minimizing the VER, but do not guarantee achieving the minimum BER and minimum SER, which are two metrics of particular importance in many applications, such as in FEC-coded systems. There are other frequently used criteria in MIMO detector design. The linear MF criterion is optimal for maximizing the received SNR in the presence of additive stochastic noise. The linear ZF criterion is optimal for maximizing the received signal-to-interference ratio (SIR). By contrast, the linear MMSE criterion is optimal for maximizing the received signal-to-interference-plus-noise ratio (SINR) amongst linear detectors [362], [363]. Additionally, the linear minimum bit-error rate (MBER) criterion based detector achieves the lowest BER amongst all linear detectors, as detailed in Section VIII-B4.

3) Computational Complexity: The optimization problem of (16) can be solved by “brute-force” search over $A^{N_I}$, resulting in an exponentially increasing computational complexity of $[A^{N_I}]$. To elaborate a little further, let us consider the example shown in Fig. 17, where binary phase-shift keying (BPSK) modulation ($M = 2$) and $N_I = 2$ are employed.

Hence, there are a total of $2^{N_I} = 4$ possible realizations for the transmitted symbol vector $s$, and they are denoted as $s_1 = [1, 1]^T$, $s_2 = [1, -1]^T$, $s_3 = [-1, -1]^T$, $s_4 = [-1, 1]^T$.

To gain deeper understanding of the computational complexity of the optimum MIMO detector formulated in (13), let us examine its implementations in practical CDMA systems. More explicitly, the optimum MUD proposed in [237] for asynchronous CDMA systems consists of a bank of MFs followed by a dynamic programming algorithm of the forward (Viterbi) type [344]–[348] (for ML criterion based detection) or of the backward-forward type [364]–[368] (for minimum error probability criterion based detection). As mentioned in Section VI, asynchronous CDMA systems can be modelled relying on the MIMO system model given in Section VI for transmission over linear dispersive channels exhibiting memory. Therefore, the optimum MUD conceived for asynchronous CDMA constitutes a sequence detector, while the optimum MUD of synchronous CDMA is a one-shot detector, and as such it is a special case of the asynchronous optimum MUD. The optimum MUD relying on brute-force search [237] requires that the transmitted energies of each user were known to the receiver. More critically, the computational complexity of the optimum decision algorithms suggested in [237], [367], [368] increases exponentially with the number of active users, i.e. it is on the order of $O(2^K)$ per bit for asynchronous transmission and $O(2^K/K)$ per bit for synchronous transmission, where $K$ is the number of active users. This is because the optimum MUD of both the synchronous and asynchronous CDMA scenarios was proven by Verdú to be an NP-hard and a non-deterministic polynomial-time complete (NP-complete) problem [46], [354]. Thus, all known algorithms designed for solving this problem optimally exhibit an exponentially increasing computational complexity in the number of decision variables. Therefore, the optimum MUD becomes computationally intractable for a large number of active users.

It should be noted that the optimal MIMO detection problem would only have a polynomially increasing complexity if and only if a polynomial-time solution could be found for any NP-complete problem, such as the famous travelling salesman problem and the integer linear programming problem which have been so far widely believed insolvable within polynomial time. However, the question of whether there exists a polynomial-time solution for NP-complete problems has not been answered by a rigorous proof to date. It is widely recognized that in computational complexity theory, the complexity class of “P” represents one of the most fundamental complexity classes, and it contains all decision problems that can be solved by a deterministic Turing machine using a polynomially increasing amount of computation time (this is conventionally abbreviated to the parlance of “polynomial-time” for convenience). In fact, the most important open question in computational complexity theory [369], [370] has been the formal proof of “Is P = NP?”, which explicitly posces the dilemma whether polynomial-time algorithms actually do exist for NP-complete problems, and by corollary, for all NP problems. Fig. 18 concisely depicts the Euler diagram characterizing the relationships amongst the P, NP, NP-complete, and NP-hard set of problems under both the $P \neq NP$ and $P=NP$ assumptions.

Additionally, it is worth mentioning that for some algorithms, such as the tree-search based MIMO detectors to be detailed in Section VIII-D, the computational complexity may

---

12In general, the MMSE detector and the linear MMSE detector are not necessarily the same. The former only aims at minimizing mean-square error (MSE) and does not impose any constraint on the form of the MMSE estimator. The latter assumes that the MMSE estimator is a linear function of the observed signal vector $y$. If $y$ and the transmitted signal vector $s$ are jointly Gaussian, then the MMSE estimator is linear. In this case, for finding the MMSE estimator, it is sufficient to find the linear MMSE estimator.

13In fact, the optimum MIMO detection problem of (16) constitutes an instance of the general closest lattice-point search (CLPS) problem, whose complexity had been analyzed earlier by Boas [45] in 1981, showing that this problem is NP-hard. Additionally, Micciancio [47] provided a simpler proof for the hardness of the CLPS problem in 2001.
vary in different scenarios. As such, the average computational complexity, the worst-case computational complexity and the distribution of computational complexity become important metrics to examine. Finally, in practical algorithm implementations, it is also important to consider the hardware complexity, which, in simplest form, can be measured by the silicon area and the number of NAND2 gates required in the IC implementation.

4) Milestone Contributions: For the sake of clarity, the main contributions to the development of the optimum MIMO detector are summarized in Table I.

The substantial performance and complexity differences between the optimum MIMO detector and the conventional SUMF detector stimulated a lot of interests in the development of suboptimum MIMO detection algorithms that are capable of achieving good performance at a low computational cost. Some representative classes of suboptimum MIMO detectors include the linear detectors, the interference cancellation aided detectors, the tree-search based detectors, the PDA based detectors, the SDPR based detectors and the LR based detectors etc., as seen in Fig. 15 and detailed below.

B. Linear MIMO Detectors

The linear MIMO detectors of Fig. 15 are based on a linear transformation of the output vector signal y. In general, they are known for their appealingly low complexity, but suffer from a considerable performance loss in comparison to the ML detector. More explicitly, the decision statistics of linear MIMO detectors may be expressed as

\[ d = Ty, \]

where \( T \) is the linear transformation (or filtering) matrix to be designed using various criteria. A conceptual illustration of the linear MIMO detectors is given in Fig. 19.

1) MF Detector: For the sake of illuminating the philosophy of linear MIMO detectors, let us rely on (4) and continue by considering the MF detector, which has the lowest computational complexity among all MIMO detectors and its linear transformation matrix is given by

\[ T_{MF} = HH^H. \]

Upon using the MF detector of (18), we obtain

\[ d = HH^Hs + HH^Hn. \]

The MF detector is well known as the optimal linear filter designed for maximizing the output SNR in the presence of additive stochastic noise. In Section VIII-A1, we have provided some discussions regarding the MF detector in order to justify the motivations of developing joint detection based MIMO detectors. To elaborate a little further, the MF detector had been widely used before the concept of MIMO detection was born, and it is essentially based on the single-user detection philosophy. Hence, strictly speaking, it does not belong to the joint detection based MIMO detection family, and typically it exhibits a poor performance in CCI-limited MIMO systems. However, in certain LS-MIMO contexts [25], [373], the MF detector is capable of approaching the performance of the optimal ML detector, as it will be further discussed in Section IX.

2) Linear ZF Detector: Assuming that the noise vector is zero, (4) becomes a system of linear equations, and the MIMO detection problem becomes equivalent to “finding the solution for \( N_t \) unknown variables subject to \( N_r \) linear equations”. Therefore, if \( H \) is a square matrix (i.e. \( N_r = N_t \)) and of full rank, the solution of this system of linear equations is given by \( s = H^{-1}y \). To generalize this problem a little further, if the matrix \( H \) satisfies \( N_r > N_t \) and has a full column rank of \( N_t \), we have \( s = H^\dagger y \), where \( H^\dagger = (HH^H)^{-1}H^H \) is the left-multiplying Moore-Penrose pseudoinverse of \( H \). This example actually conveys the essential idea of the ZF criterion based MIMO detector, for which the linear transformation matrix is given by

\[ T_{ZF} = H^\dagger, \]

and if \( H \) is invertible, the left-multiplying pseudoinverse \( H^\dagger \) and the inverse coincides, i.e. we have \( H^\dagger = H^{-1} \). Upon using the ZF detector, we have \( d = s + H^\dagger n \), which indicates that the interference amongst the multiple inputs is completely eliminated, albeit the noise power is augmented.

Similar to the case of the optimum ML-based MIMO detector, the ZF criterion based linear MIMO detector of Fig. 15 was also first proposed by van Etten [228] in 1975 for a multiple-channel multiplexing transmission system subjected to both ISI and ICI. As far as CDMA systems are concerned,
TABLE I
MILESTONES IN THE DEVELOPMENT OF THE OPTIMAL MIMO DETECTOR

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1976</td>
<td>van Ettan [229]</td>
<td>Derived an ML sequence estimation based receiver for combating both the ISI and ICI in multiple-channel transmission systems and demonstrated that under certain conditions, the performance of the ML receiver is asymptotically as good as if both the ISI and ICI were absent.</td>
</tr>
<tr>
<td>1981</td>
<td>Boas [45]</td>
<td>Analyzed the complexity of the generic problem of &quot;closest point search in an N_f-dimensional lattice&quot;, which is identical to the optimum MIMO detection problem, as a function of the dimension N_f of the decision-variable vector, and proved that this problem is NP-hard. Thus, all known algorithms conceived for solving the generic MIMO detection problem optimally have an exponentially increasing computational complexity.</td>
</tr>
<tr>
<td>1983 - 1986</td>
<td>Verdú [235]-[237], [354]</td>
<td>First presented a full derivation and analysis of the ML based multiuser detector for asynchronous/synchronous CDMA systems; showed that there is, in general, a huge gap between the performance of the conventional SUMF and the optimal attainable performance; showed that the infamous near-far problem was not an inherent flaw of CDMA but a consequence of the inability of the SUMF to exploit the structure of the MUI; introduced the performance measure of multiuser asymptotic efficiency, which was later widely used in the asymptotic analysis of multiuser detectors at the high-SNR region.</td>
</tr>
<tr>
<td>1984 - 1989</td>
<td>Verdú [46], [354]</td>
<td>Independently proved that the optimum MUD problem in CDMA systems is NP-hard and NP-complete.</td>
</tr>
<tr>
<td>2001</td>
<td>Micciancio [47]</td>
<td>Presented a simpler proof of the NP-hardness of the problem of closest point search in an N_f-dimensional lattice.</td>
</tr>
<tr>
<td>2003</td>
<td>Garrett et al. [371]</td>
<td>Proposed the first VLSI implementation of a soft-output ML detector having a 19.2 Mbps uncoded data rate supporting up to 4 × 4 QPSK MIMO.</td>
</tr>
<tr>
<td>2003</td>
<td>Burg et al. [372]</td>
<td>Presented an efficient VLSI implementation of hard-decision optimum ML detector for QPSK MIMO. The proposed method does not compromise optimality of the ML detector. Instead it uses the special properties of QPSK modulation, together with algebraic transformations and architectural optimizations, to achieve low hardware complexity and high speed up to 50 Mbps.</td>
</tr>
</tbody>
</table>

this solution was first proposed by Schneider [231] in 1979 for synchronous CDMA systems transmitting equal-energy multiuser signals, where he sought to minimize the probability of bit error, but erroneously arrived at the ZF detector. From 1986 to 1990, Lupas and Verdú systematically investigated this detector in the context of both synchronous [239], [374] and asynchronous [240], [375] CDMA systems. They referred to it as the linear decorrelating multiuser detector. It was shown that if the transmitted energies of each user are unknown to the receiver, then both the ML amplitude estimates and the ML decisions on the transmitted bits are obtained by the ZF detector, regardless of the values of the received energies of each user. As a beneficial result, the ZF detector achieves the same degree of resistance to the infamous near-far problem as the optimum ML detector, despite its significantly reduced computational complexity. The insight that the near-far problem was not an inherent flaw of CDMA but a consequence of the SUMF’s inability to exploit the non-Gaussian structure of the MUI [237], and the fact that the joint detection based MUDs, including its linear versions, achieve a significantly better near-far resistance [239], [240], [374], [375] became another major incentive for the subsequent research activities dedicated to MUD in CDMA. Additionally, with the advent of the multiple-antenna technologies conceived during the mid-1990s, the ZF detector was first studied in the SDM-based VBLAST systems by Foschini, Wolniansky, Golden and Valenzuela [179]-[181].

3) Linear MMSE Detector: As seen in Fig. 15, the linear transformation matrix $T$ of (17) can also be designed according to the MMSE criterion, which minimizes the mean-square error between the actual transmitted data and the channel’s output data after using the linear transformation matrix $T$. To be more specific, $T$ is obtained by solving the optimization problem of

$$
T_{\text{MMSE}} = \arg \min_{T} \mathcal{E} \left( \| s - Ty \|_2^2 \right). \tag{21}
$$

Using the orthogonality principle [376], we have

$$
\mathcal{E}(s - Ty) = 0, \tag{22}
$$

then $T_{\text{MMSE}}$ may be derived as

$$
T_{\text{MMSE}} = (H^TH + 2\sigma^2 I)^{-1}H^H, \tag{23}
$$

where $\sigma^2$ is the noise power per real dimension, and $\mathcal{E}(s) = 1$ is assumed. Compared to the linear ZF detector, the linear MMSE detector achieves a better balance between the MUI elimination and noise enhancement by jointly minimizing the total error imposed by both the MUI and the noise. Hence, the linear MMSE detector achieves a better performance at low SNRs than the ZF detector.

The MMSE criterion based linear MIMO detector was first proposed by Shnidman [226] in 1967, and hence it is the oldest MIMO detector found in the literature. The generalized Nyquist criterion formulated by Shnidman first
indicates that the ISI and crosstalk\textsuperscript{14} between multiplexed signals essentially represent identical phenomena. Then, relying on this insight, he proposed a linear receiver that is optimal in the sense of the MMSE criterion for combating both the ISI and crosstalk in single-channel multiple-waveform-multiplexed pulse-amplitude modulation (PAM) systems. In 1970, Kaye and George [227] explicitly extended the MMSE receiver of [226] to the family of general multiple-channel systems transmitting multiplexed PAM signals and/or providing diversity. The MMSE criterion based linear detector for CDMA systems was proposed by Xie, Rushforth and Short in 1989 [253], [377]. A decade later, it was also revisited by Foschini, Wolniansky, Golden and Valenzuela in the context of SDM-based multi-antenna systems [179]–[181]. The performance of linear ZF/MMSE based MIMO detectors depends on the SINR experienced at the output of these detectors, which was first analyzed by Poor and Verdú [378] in 1997, and investigated in more depth later from various other perspectives, such as the error probability, outage probability, diversity-multiplexing tradeoff (DMT) [379], [380], as well as asymptotic distribution of the SINR (in terms of antenna number and high/low SNR regimes) [362], [363], [381]–[385].

4) Other Linear Detectors: As observed in the family-tree of Fig. 15, there are a range of other criteria for designing the linear transformation matrix $T$.

- For example, in [239], [374], Lupas and Verdú also proposed a maximum asymptotic-multiuser-efficiency (MAME) based linear detector, which is capable of minimizing the probability of bit errors in the limit as the noise approaches zero. The asymptotic-multiuser-efficiency (AME) is a metric which characterizes the performance of the MUD in the high-SNR region. It implies the performance loss of the desired user in the high-SNR region due to the interference imposed by other active users. To be more specific, it is defined as the limit of the ratio between the effective SNR (that is required by a single-user system to achieve the same asymptotic error probability) and the actual SNR of the desired user, when the noise power tends to zero. Furthermore, the linear MAME detector was designed by exploiting the assumption that the individual transmitted energies of all the users are fixed and known to the receiver. By contrast, the ZF detector does not require the knowledge of the transmitted energies of the users.

- Additionally, since a common disadvantage of the linear ZF and MMSE detectors is that their estimates of the transmitted symbols are biased, Xie, Rushforth and Short [253], [377] proposed the so-called weighted least-squares (WLS) linear detector, which is capable of providing an unbiased estimate of the transmitted symbols. It is worth pointing out that except for the linear MF and ZF detectors, other linear MIMO detectors – including the linear MMSE detector, the linear MAME detector and the linear WLS detector – were typically derived under the assumption that the system parameters such as the signal’s phase, power and delay are known. As a result, in practice these parameters must be estimated and the receiver’s structure has to be regularly modified to reflect the updated estimates.

- Another important class of linear MIMO detectors are based on the MBER criterion. The linear MBER detector is capable of outperforming the linear MMSE detector when either the signature cross-correlation is high or the background noise is non-Gaussian [392]. Again, the MBER based MIMO detector was first considered by van Etten [228] in 1975 in the context of a multiple-channel multiplexing transmission system subjected to both ISI and ICI. This MBER criterion was later studied in the context of CDMA systems [386]–[389], [392]–[397] and multi-antenna systems [308], [309].

- Finally, we would like to mention that the linear MIMO detector can also be designed from the perspective of a linear equalizer [253], [377], since the mathematical models of the MIMO detection problem and of the equalization problem are similar [343]. To elaborate a little further, in MIMO systems each symbol’s interference is imposed by other simultaneous transmissions, while in the band-limited ISI channels requiring equalization, the interference of a particular symbol is due to other symbols that are transmitted sequentially in the time domain.

The main contributions to the development of linear MIMO detectors are summarized in Table II.

C. Interference Cancellation Aided MIMO Detectors

Another important class of suboptimum MIMO detectors portrayed in Fig. 15 are constituted by the interference cancellation based MIMO detectors, which are nonlinear and generally achieve a better performance than linear MIMO detectors. The concept of interference cancellation was first studied in 1974 by Bergmans and Cover [398], [399], as well as by Carleial [197] in 1975, in their information-theoretic studies of broadcast channels and of interference channels, respectively. In the context of CDMA and multi-antenna systems, this class of MIMO detectors have numerous variants due to the associated design flexibility, including the successive interference cancellation (SIC) detector [179]–[181], [251], [400], the parallel interference cancellation (PIC) detector [241], [248], [318], the multistage interference cancellation (MIC) detector [234], [244], [245], [401], and the decision-feedback detector (DFD) [253], [255], [256], [258] et al. The interference cancellation based MIMO detectors are typically capable of providing a significantly better performance than their linear counterparts at the expense of a higher complexity, especially in the absence of channel coding [260], albeit this is not necessarily always the case. In practice, a common drawback of the interference cancellation based MIMO detectors is that they often suffer from error propagation. Hence their performance only approaches that of the optimum ML based MUD when the interfering users have a much stronger signal strength than the desired user. From this perspective,
the weakest user benefits most from the employment of the interference cancellation detector.

- **SIC**: In the most popular SIC based MIMO detector, a single symbol $s_k$ is detected at a time. Then the interference imposed by this particular symbol on the other symbols $\{s_{k'}\}_{k' \neq k}$ yet to be detected is subtracted after recreating the interference upon generating the modulated signal corresponding to this symbol. In this scheme, it is most important to cancel the effect of the strongest interfering signal before detecting the weaker signals. Therefore, the specific symbol detection ordering, which can be designed according to various criteria, is quite critical for the SIC detector’s performance. Some of the typical ordering criteria for ordered SIC (OSIC) include the decreasing signal-to-noise ratio (DSNR) criterion [310], [311], the greatest signal-to-noise ratio (GSNR) criterion [312], the increasing mean-square error (IMSE) criterion [311], and the least mean-square error (LMSE) criterion [313]–[315]. The SIC method performs well when there is a substantial difference in the received signal strength of the multiple simultaneously transmitted symbols. However, this condition is not always satisfied in practical applications, which renders the SIC detector potentially sensitive to decision error propagation. Therefore, the SIC detector is well-suited for multiple-access systems suffering from the near-far problem, such as the

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1967</td>
<td>Shnidman [226]</td>
<td>First formulated a generalized Nyquist criterion, which pointed out that the ISI and crosstalk between multiplexed signals are essentially identical phenomena; he then proposed a linear MMSE receiver for combating both ISI and crosstalk in single-channel multiple-waveform multiplexed systems.</td>
</tr>
<tr>
<td>1970</td>
<td>Kaye et al. [227]</td>
<td>Extended the MMSE receiver of [226] to the general multiple-channel systems transmitting multiplexed PAM signals and/or providing diversity.</td>
</tr>
<tr>
<td>1975</td>
<td>van Etten [228]</td>
<td>Developed linear receivers based on both the ZF criterion and the minimum error probability criterion for a multiple-channel transmission system similar to that of [227]; these two detectors heralded the linear ZF and the linear MBER multiuser detectors of CDMA systems.</td>
</tr>
<tr>
<td>1975</td>
<td>Horwood et al. [230]</td>
<td>Proposed two linear signal-correlation based detectors for synchronous digital multiple-access systems; one of them assumes that each user only knows its own signature, while the other assumes that each user knows all users’ signatures; this is the first attempt in multiple-access systems to exploit the structure of the signals simultaneously sent, which is the key idea of MUD in CDMA systems.</td>
</tr>
<tr>
<td>1979</td>
<td>Schneider [231]</td>
<td>First made an attempt to conceive MUD for CDMA systems; he proposed the linear decorrelating detector, namely the linear ZF detector, which represents one of the mainstream MUD approaches conceived for CDMA systems; this detector was also extended to the scenario of combating crosstalk in M-ary multiplexed transmission systems in 1980 [232].</td>
</tr>
<tr>
<td>1986-1990</td>
<td>Lupas et al. [239], [240], [374], [375]</td>
<td>Systematically investigated the linear ZF MUD in the context of both synchronous [239], [374] and asynchronous [240], [375] CDMA systems; they showed that the ZF detector achieves exactly the same degree of resistance to the infamous near-far problem as the optimum ML detector, despite its much lower computational and implementation complexity; they also first proposed a linear MAME MUD, which is capable of equivalently minimizing the probability of bit error in the limit as the noise approaches zero.</td>
</tr>
<tr>
<td>1989 - 1990</td>
<td>Xie et al. [253], [377]</td>
<td>First proposed the MMSE criterion based linear MUD, the modified linear equalizer based MUD, and the WLS linear MUD for CDMA systems. In contrast to the linear ZF detector, to the linear MMSE detector, and to the modified linear equalizer based detector, the linear WLS detector is capable of providing an unbiased estimate of the transmitted symbols.</td>
</tr>
<tr>
<td>1993 - 1997</td>
<td>Mandayam et al. [386]–[389]</td>
<td>First proposed the MBER criterion based linear MIMO detectors for CDMA systems; the linear MBER detector is capable of outperforming the linear MMSE detector when either the signature cross-correlation is high or the background noise is non-Gaussian.</td>
</tr>
<tr>
<td>2006</td>
<td>Chen et al. [308]</td>
<td>Proposed the MBER criterion based linear detector for multi-antenna aided MIMO systems.</td>
</tr>
<tr>
<td>2006</td>
<td>Burg et al. [390]</td>
<td>Presented an algorithm and a corresponding VLSI architecture for the implementation of linear MMSE detection in packet-based MIMO-OFDM communication systems. The algorithm also supports the extraction of soft information for channel decoding.</td>
</tr>
<tr>
<td>2009</td>
<td>Yoshizawa et al. [391]</td>
<td>Reported a VLSI implementation for a 4 × 4 MIMO-OFDM transceiver relying on linear MMSE, which achieves a target data rate of 1 Gbps.</td>
</tr>
<tr>
<td>2014</td>
<td>Yin et al. [341]</td>
<td>Presented the first application-specific integrated circuit (ASIC) implementation for the soft-output linear MMSE detector based large-scale MIMO system which uses 128 BS antennas to support 8 users, and a sum-rate of 3.8 Gbps was achieved.</td>
</tr>
</tbody>
</table>
family of CDMA or SDMA systems. In the SIC detector, there is a need for detection reordering at each iteration of the SIC detector, and the number of detection iterations increases linearly with the number of symbols in \( s \). Therefore, for a system which has a high-dimensional transmitted symbol vector \( s \), the SIC technique imposes a substantial complexity, which ultimately increases the processing delay. For the SIC detector designed for CDMA systems was first proposed by Viterbi [251]. Later it was studied extensively in [257], [259], [400], [402]–[407]. In the context of multi-antenna based SDM systems, the SIC scheme was first studied by Foschini, Wolniansky, Golden and Valenzuela in [179]–[181], and it was later studied more comprehensively by numerous other researchers in [305]–[307]. Among these schemes, Viterbi [251] proposed an SIC scheme for a convolutionally coded direct-sequence CDMA (DS-CDMA) system and revealed that with the aid of the SIC based receiver, the aggregate data rate of all simultaneous users may approach the Shannon capacity of the Gaussian channel. It should be emphasized that although theoretically the SIC method achieves the Shannon capacity in the multiple-access channel by assuming perfectly error-free detection (hence avoiding decision error propagation), this is not necessarily true in practice, because the SIC method is sensitive to decision error propagation, and hence MIMO detectors that are more robust to decision error propagation might outperform the SIC detector in practice. Another fact worth mentioning is that the performance degradation imposed by error propagation in the SIC detector can be mitigated by accurate power control [408].

- **PIC**: Alternatively, in the PIC based MIMO detector, all symbols are detected simultaneously. For each symbol, the coarse initial estimate of the interfering symbols can be used for regenerating the interference and then for deducting it from each of the composite received signals. Then this PIC detection process may be repeated for several iterations. Therefore, sometimes the PIC detection is also regarded as a MIC technique, or vice versa. Compared to the SIC detector, the PIC detector has lower processing delay, and is more robust to inter-stream error propagation. However, its near-far resistance is inferior to that of the SIC detector, because some users might have much weaker received signal strength than others. Hence, the PIC is suitable for similar-power signals, while the SIC performs better for different-power streams. In the context of CDMA systems, the earliest contribution to PIC may be attributed to Kohno et al. [241]–[243]. Later significant contributions to PIC were also attributed to Yoon [247], [409], Divsalar [248], Buehrer [410] and Guo [411] et al.. In the context of multi-antenna MIMO systems, the PIC detector was studied mainly in [316]–[318].

- **MIC**: In the MIC based MIMO detector, the first stage can be the conventional SUMF detector, the linear ZF/MMSE detector, the SIC detector or any other suboptimum detector. The decisions made for all symbols \( s \) by the \((n-1)\)th stage are employed as the input of the \( n \)th stage for the sake of cancelling the MUI. Note that historically, the MIC detector was developed independently of the PIC, although they share similar concepts. The MIC detector was first proposed by Timor for frequency-hopped CDMA (FH-CDMA) systems [233], [234]. It was then extensively studied in the context of both asynchronous DS-CDMA systems [244], [412] and synchronous DS-CDMA systems [245], [401], [413]. An analytical framework was proposed for adaptive MIC in [249].

- **DFD**: The concept of DFD is based on the same premise as that of the family of decision-feedback equalizers [417], [418]. Although DFD also relies on the SIC idea, its emphasis is on the receiver filter’s optimization, which consists of a feedforward filter and a feedback filter optimization. The first DFD scheme was proposed by Xie et al. [253], [377] for asynchronous DS-CDMA systems. Other major contributors of the subject of DFD include Duel-Hallen [255], [256], [414], [415], who comprehensively investigated decision-feedback MUDs designed for both synchronous [255], [414] and asynchronous [256], [415] CDMA systems. Furthermore, Varanasi [258] analyzed the performance of a general class of DFDs using a new performance metric constituted by the probability that at least one user is detected erroneously, and also proposed algorithms for determining the most beneficial detection ordering.

The basic principles of the SIC/DFD detectors and the PIC/MIC detectors are illustrated in Fig. 20 and Fig. 21, respectively. The main contributions to the development of the interference cancellation based MIMO detectors are summarized in Table III. A more comprehensive exposition of the above-mentioned MIMO detectors developed in the context of CDMA systems can be found in [48]–[51], [261], [419].

D. Tree-Search Based MIMO Detectors

The tree-search based MIMO detectors are arguably the most popular detectors investigated in the era of multi-antenna
MIMO systems. This is because 1) the introduction of the powerful SD algorithm happened to coincide with the development of multi-antenna MIMO techniques; 2) some profound research results on the CLPS problem showed that the tree-search MIMO detectors enjoy significant design flexibility in terms of striking an attractive tradeoff between approaching the optimum ML performance and reducing the computational complexity.

Indeed, some tree-search based MUDs had been reported earlier in the context of CDMA systems [252], [254], [420]–[422]. For example, the so-called (depth-first) stack sequential detection was proposed by Xie in [252], [420], while the (breadth-first) K-best tree-search detection was proposed, again, by Xie in [254], [421], which was then further studied by Wei et al. in [422]. Looking back to the earlier history, because of the convertibility between the trellis structure and the tree structure, the tree-search detection methods proposed for CDMA systems, including the classic M-algorithm [423], [424] and T-algorithm [425]–[428], were actually extensions of their counterparts used in trellis decoding [423]–[436]. However, these tree-search based detectors did not attract as much attention as the linear detectors and the interference cancellation aided detectors in the era of CDMA systems.

The research interests related to tree-search based MIMO detectors were largely stimulated by the seminal work of Viterbo et al. [53], [319], who first applied the depth-first SD algorithm to the ML detection of multidimensional constellations transmitted over single-antenna fading channels. The basic principle of the SD algorithm is illustrated in Fig. 22. Compared to the optimal brute-force search based ML decoding, the SD algorithm aims for reducing the computational complexity by only searching over the noiseless received signals that lie within a hypersphere of radius \( R \) around the received signal. Note that before it was applied in digital

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1974-1975</td>
<td>Bergmans and Cover [398], [399]</td>
<td>First demonstrated the effectiveness of the SIC concept from an information-theoretic perspective for broadcast channels.</td>
</tr>
<tr>
<td>1975</td>
<td>Carleial [197]</td>
<td>First characterized the effectiveness of the SIC principle from an information-theoretic perspective for interference channels.</td>
</tr>
<tr>
<td>1980-1981</td>
<td>Timor [233], [234]</td>
<td>First proposed a two-stage [233] MUD and a multistage [234] MUD for FH-CDMA systems employing multiple frequency-shift keying (MFSK) modulation; showed that the mutual interference between the users of a FH-CDMA system may be significantly reduced by making use of the well-defined algebraic structure of the users’ signature waveforms, and that introducing an extra stage of interference cancellation may further improve the detector’s performance.</td>
</tr>
<tr>
<td>1990</td>
<td>Viterbi [251]</td>
<td>First conceived an SIC scheme for a convolutionally coded DS-CDMA system, and revealed that with the aid of the SIC based receiver the aggregate data rate of all simultaneous users can approach the Shannon capacity of the Gaussian channel.</td>
</tr>
<tr>
<td>1993-1999</td>
<td>Kohno et al. [241]–[243]</td>
<td>First proposed a PIC based MUD for CDMA systems.</td>
</tr>
<tr>
<td>1988-1991</td>
<td>Varanasi et al. [244], [245], [412], [413]</td>
<td>Designed and systematically characterized the MIC MUDs for both asynchronous and synchronous CDMA systems.</td>
</tr>
<tr>
<td>1989-1990</td>
<td>Xie et al. [253], [377]</td>
<td>First proposed a DFD based MUD for asynchronous DS-CDMA systems.</td>
</tr>
<tr>
<td>1991-1995</td>
<td>Duel-Hallen et al. [255], [256], [414], [415]</td>
<td>Systematically investigated DFD MUDs conceived for both synchronous [255], [414] and asynchronous [256], [415] CDMA systems from a receiver filter optimization perspective.</td>
</tr>
<tr>
<td>2002</td>
<td>Chin et al. [316]</td>
<td>Extended the PIC detector to the multiple-antenna aided SDM-MIMO systems.</td>
</tr>
<tr>
<td>2003</td>
<td>Wübben et al. [310]</td>
<td>Proposed a QR-decomposition (QRD) based MMSE-SIC detector for multiple-antenna aided SDM-MIMO systems.</td>
</tr>
<tr>
<td>2003</td>
<td>Guo et al. [416]</td>
<td>Presented a VLSI implementation of the V-BLAST detector for a 4 × 4 MIMO system employing QPSK, and a detecting throughput of 128 Mbps was achieved.</td>
</tr>
<tr>
<td>2011</td>
<td>Studer et al. [318]</td>
<td>Reported an ASIC implementation of a soft-input soft-output MMSE based PIC detector for multiple-antenna aided SDM-MIMO systems.</td>
</tr>
</tbody>
</table>
communications, the SD algorithm, also known as the Fincke-Pohst algorithm, had been reported in [437], [438]. Later, Agrell et al. [54] proposed to employ the Schnorr-Euchner (SE) refinement [439] of the Fincke-Pohst algorithm [437], [438] for solving the CLPS problem, and they concluded that the SE enumeration is more efficient than the Viterbo-Boutros (VB) implementation [53] of the SD algorithm. More recently, Damen et al. [57] proposed two improved SD algorithms for finding the closest lattice point, both of which were shown to offer a significant complexity reduction compared to the VB-SD [53] and to the SE-SD [54], respectively. There exist a number of other variants of the tree-search based MIMO detectors, which mainly fall into three categories: the depth-first tree-search detector [53], [54], [57]–[61], [80], [82], [319], the breadth-first tree-search detector [62]–[64], [320] and the best-first tree-search detector [68]–[72], [321], [322], [440].

The major momentum which propels the enormous research activities on tree-search based MIMO detectors is that they were shown to be capable of achieving near-ML performance, or even exact ML performance at the expense of significantly reduced complexity [58], [59], [441]–[443]. However, we would like to emphasize that this claim is shown not true in general [60], [444]. More specifically, Hassibi and Vikalo [58], [59], [441]–[443] first studied the expected complexity, averaged over the noise and over the lattice, of the Fincke-Pohst SD based MIMO detectors. It was claimed that although the worst-case complexity of the SD algorithm is exponential, the expected complexity of the SD algorithm is polynomial, in fact, often roughly cubic, for a wide range of SNRs and number of antennas. Contrary to this claim, Jaldén and Ottersten [60], [444] demonstrated that the expected complexity of SD based MIMO detectors is given by \( O(M^{3N_t}) \), where \( \beta \in (0,1) \) is a small factor depending on the value of SNR. In other words, the expected complexity of the SD algorithm is still exponential for fixed SNR values. Therefore, in general the tree-search based MIMO detectors are not efficient for MIMO systems which operate under low-SNR condition and/or have a large number of inputs. Notably, in order to avoid the varying-complexity characteristic of tree-search based MIMO detectors, recently a suboptimal fixed-complexity SD (FCSD) was proposed for MIMO systems [80]. It was shown that the FCSD achieves a near-ML performance with a complexity of \( O(M^{3\sqrt{N_t}}) \) [82] regardless of the specific SNR encountered, which represents an attractive solution of facilitating an efficient hardware implementation compared to the conventional exponential-complexity SD. The main contributions in the development of the depth-first tree-search MIMO detectors, the breadth-first tree-search MIMO detectors and the best-first tree-search MIMO detectors are summarized in Table IV, Table V and Table VI, respectively.

### E. Lattice-Reduction Aided Detectors

Lattice-reduction (LR) aided detectors constitute another important class of MIMO detectors, which rely on the algebraic concept of “lattice” originating from classic geometry and group theory. A lattice in \( \mathbb{R}^n \) is a discrete subgroup of \( \mathbb{R}^n \), which spans the real-valued vector space \( \mathbb{R}^n \). Each lattice in \( \mathbb{R}^n \) can be generated from a basis of the vector space by forming all linear combinations with integer coefficients. In the MIMO transmission model of (1), the received signal vector \( y \) is the noisy observation of the vector \( Hs \), which is in the lattice spanned by the column vectors \( H \), since both the real and imaginary parts of all the elements of \( s \) may be transformed to integers by shifting and scaling.

A lattice typically has multiple sets of basis vectors. Some bases spanning the same lattice as \( H \) might be “closer” to orthogonality than \( H \) itself. The process of finding a basis closer to orthogonality is referred to as LR. Theoretically, finding the optimal (closest to orthogonality) set of basis vectors is computationally expensive. Therefore, in practice LR algorithms typically aim for finding a “better” channel matrix \( H = HL \), where the real and imaginary parts of all the entries of the unimodular matrix \( L \) are integers and the determinant of \( L \) is \( \pm 1 \) or \( \pm j \). As a result, the LR preprocessing technique is capable of improving the “quality” of the MIMO channel matrices.

There is a variety of LR algorithms developed by mathematicians [454]. Some of them, such as Gaussian reduction [455], Minkowski reduction and Korkine-Zolotareff (KZ) reduction [456], are capable of finding the optimal basis for a lattice, but they are computationally prohibitive for communications systems [456]–[459]. Other well-known LR algorithms include the Lenstra-Lenstra-Lovász (LLL) algorithm [460], Seysen’s algorithm [90], [91], [461] and Brun’s algorithm [92], [462], [463], which are all suboptimal. The most popular LR algorithm is the LLL algorithm, which does not guarantee to find the optimal basis, but it guarantees to find a basis within a factor to the optimal one in polynomial time [88], [89], [457], [464]. For example, it was formally proved in [88] that an upper bound on the average computational complexity of LLL is \( O(N_t N_r \log N_t) \), where \( N_t \) is the size of \( s \). Furthermore, a tighter upper bound of \( O(N_t^2 \log N_t N_r) \) was found in [89], where \( N_r \) is the size of \( y \). Note that the worst-case computational complexity of the LLL algorithm can be infinite [89], [457]. However, in practice the probability of the scenario which leads to this worst-case complexity is zero. There are mainly two variants for the LLL algorithm, namely the real-valued LLL [84], [85] and the complex-valued LLL [95]–[98]. The real-valued LLL is applied to the real-valued MIMO system model of Section VII, while the complex-valued LLL is designed for directly using it in the complex-valued MIMO system model. Additionally, the authors of [465] proposed an LLL algorithm which is not only applicable to the complex-valued model, but also applicable to the Euclidean ring in general.

In principle, LR can be combined with virtually all the other MIMO detectors to further improve their performance. For example, the LR technique was used in conjunction with traditional linear ZF and nonlinear ZF-SIC detectors in [83], as

\[ \text{Recall that the lattice perspective – many detection problems can be interpreted as the problem of finding the closest lattice point} \]


<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981 - 1985</td>
<td>Pohst and Fincke [437], [438]</td>
<td>First proposed the SD algorithm, which is hence known as the Fincke-Pohst algorithm, for calculating vectors of short length in a lattice at a reduced complexity; this work laid the mathematical foundation of applying the SD algorithm to the MIMO detection problem.</td>
</tr>
<tr>
<td>1988 - 1990</td>
<td>Xie et al. [252], [420]</td>
<td>First proposed a stack sequential decoding based MUD for asynchronous CDMA systems; this detector is essentially a depth-first tree-search MIMO detector.</td>
</tr>
<tr>
<td>1993 - 1999</td>
<td>Viterbo et al. [53], [319]</td>
<td>Applied the depth-first SD algorithm to the ML detection of multidimensional constellations transmitted over single-antenna fading channels, which largely stimulated the research interests of tree-search based MIMO detectors.</td>
</tr>
<tr>
<td>1994</td>
<td>Schnorr and Euchner [439]</td>
<td>Proposed a more efficient variation, known as SE refinement, of the Fincke-Pohst SD algorithm; the SE-SD algorithm was based on the lattice basis reduction philosophy and represents a popular solution to the MIMO detection problem.</td>
</tr>
<tr>
<td>2001 - 2003</td>
<td>Hochwald et al. [74], [445]</td>
<td>Proposed a complex-valued SD and the list-SD (LSD) for a FEC-coded MIMO using IDD receiver, showing that a near-capacity performance can be achieved with the aid of a soft-SD based IDD receiver.</td>
</tr>
<tr>
<td>2002</td>
<td>Agrell et al. [54]</td>
<td>First proposed to use the SE refinement [439] of the Fincke-Pohst SD algorithm [437], [438] to the CLPS problem, and concluded that the SE enumeration technique is more efficient than the VB implementation [53] of the SD algorithm designed for MIMO detection.</td>
</tr>
<tr>
<td>2003</td>
<td>Damen et al. [57]</td>
<td>Proposed a pair of improved SD algorithms for finding the closest lattice point, both of which were shown to offer a significant complexity reduction compared to the VB-SD of [53] and to the SE-SD of [54].</td>
</tr>
<tr>
<td>2001 - 2005</td>
<td>Hassibi and Vikalo [58], [59], [441]-[443]</td>
<td>Analyzed the expected complexity of the SD algorithm, and concluded that the expected complexity of SD algorithm is dependent on both the problem size and the SNR; showed that when the SNR is high, the expected complexity of SD can be approximated by a polynomial function for a small problem size.</td>
</tr>
<tr>
<td>2004 - 2005</td>
<td>Jaldén and Ottersten [60], [444]</td>
<td>Further analyzed the expected complexity of the SD algorithm, and demonstrated that the expected complexity of the SD algorithm increases exponentially for a fixed SNR with a search-space, which contradicts previous claims; therefore, strictly speaking, the SD algorithm has an exponential lower bound in terms of both the expected complexity as well as the worst-case complexity, although it can be efficient at high SNRs and for problems of moderate size.</td>
</tr>
<tr>
<td>2004</td>
<td>Garrett et al. [446]</td>
<td>First reported the VLSI implementation of a soft-output depth-first SD based detector for a 4 × 4 16QAM MIMO system, achieving 38.8 Mbps over a 5-MHz channel.</td>
</tr>
<tr>
<td>2005</td>
<td>Burg et al. [61]</td>
<td>Proposed two ASIC implementations of depth-first MIMO SD. The first ASIC attains the ML performance with an average throughput of 73 Mbs/s at an SNR of 20 dB; the second ASIC achieves a throughput of 170 Mbs at the same SNR with only a negligible BER degradation. The proposed implementations rely on four key contributing factors to achieve high throughput and low complexity: depth-first tree traversal with radius reduction, implemented in a one-node-per-cycle architecture, the use of the $l^\infty$-norm, and an efficient implementation of the enumeration approach.</td>
</tr>
<tr>
<td>2006 - 2008</td>
<td>Barbero et al. [80], [447]</td>
<td>Proposed a noise-level independent fixed-complexity tree-search MIMO detector, which overcomes the two main limitations of the SD from an implementation point of view: its variable complexity and its sequential nature.</td>
</tr>
<tr>
<td>2008</td>
<td>Studer et al. [65]</td>
<td>Presented the VLSI implementation of a soft-output depth-first SD based MIMO detector, which demonstrated that single tree-search, sorted QR-decomposition, channel matrix regularization, log-likelihood ratio clipping, and imposing runtime constraints are the key ingredients for realizing soft-output MIMO detectors with near max-log performance.</td>
</tr>
<tr>
<td>2009</td>
<td>Jaldén et al. [82]</td>
<td>Presented analytical study of the error probability of the fixed-complexity SD in MIMO systems having an arbitrary number of antennas, proving that it achieves the same diversity order as the ML detector, regardless of the constellation size used.</td>
</tr>
</tbody>
</table>
well as with linear MMSE and nonlinear MMSE-SIC detectors in [85], [466], both achieving a substantial performance gain with little additional computational complexity. As a further advance over [83], a real-valued LLL-based LR algorithm was used in [84], which enables the application of the algorithm in MIMO systems with arbitrary numbers of dimensions. In addition, it was shown in [84], [86] that LR can also be favorably applied in MIMO systems that use precoding. The LLL-based LR algorithm was shown to be capable of achieving the maximum attainable/full receive diversity in MIMO decoding [87]. The VLSI implementation of the LR technique aided precoder and of the $K$-best MIMO detector was reported in [92] and [93], respectively. LR-aided soft-decision MIMO detectors are studied in [99]–[102]. Recently, element-based LR algorithms, which reduce the diagonal elements of the noise covariance matrix of linear detectors and thus enhance the asymptotic performance of linear detectors, were proposed for large-scale MIMO systems [103]. The main contributions in the development of LR-aided MIMO detection are summarized in Table VII.

### F. Probabilistic Data Association Based Detectors

The PDA filter technique is a statistical approach originally invented by Bar-Shalom [467] in 1975 for the problem of

---

### TABLE V

**MILESTONES IN THE DEVELOPMENT OF THE TREE-SEARCH MIMO DETECTORS: BREADTH-FIRST TYPE**

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990 - 1993</td>
<td>Xie et al. [254], [421], [422]</td>
<td>First conceived a breadth-first $K$-best tree search MUD for asynchronous CDMA systems; proposed a joint signal detection and parameter estimation scheme based on their breadth-first tree search MUD.</td>
</tr>
<tr>
<td>1997</td>
<td>Wei et al. [422]</td>
<td>Studied both the $M$-algorithm and the $T$-algorithm based breath-first tree-search MUD in the context of CDMA systems operating in fading channels.</td>
</tr>
<tr>
<td>2002</td>
<td>Wong et al. [62]</td>
<td>Proposed and implemented a breadth-first $K$-best tree-search MIMO detector using a VLSI architecture, which is capable of achieving a decoding throughput of 10 Mb/s at 100 MHz clock frequency in a 16-QAM aided $(4 \times 4)$-element SDM-MIMO system.</td>
</tr>
<tr>
<td>2004 - 2006</td>
<td>Guo et al. [63], [448]</td>
<td>Proposed and implemented both hard and soft SE-strategy based $K$-best tree-search MIMO detectors, which are capable of supporting up to 53.3 Mb/s throughput at 100 MHz clock frequency for a 16-QAM aided $(4 \times 4)$-element SDM-MIMO system.</td>
</tr>
<tr>
<td>2006</td>
<td>Wenk et al. [449]</td>
<td>Presented a new VLSI architecture for the implementation of the $K$-best algorithm, which relies on a more parallel approach and the ASIC design achieves up to 424 Mbps throughput.</td>
</tr>
<tr>
<td>2007</td>
<td>Chen et al. [64]</td>
<td>Reported a VLSI implementation of a soft-output breadth-first tree search aided MIMO detector for a $(4 \times 4)$-element MIMO system employing 64-QAM, which is capable of achieving a throughput of above 100 Mb/s.</td>
</tr>
<tr>
<td>2010</td>
<td>Patel et al. [450]</td>
<td>Presented a VLSI architecture of a novel soft-output K-Best MIMO detector. This implementation attains a peak throughput of 635 Mbps for a $4 \times 4$ 64-QAM MIMO system with 0.13um CMOS. Synthesis results in 65nm CMOS show the potential to support a sustained throughput up to 2 Gbps, which may meet the requirements of for mobile WiMAX and LTE-A standards.</td>
</tr>
</tbody>
</table>

---

### TABLE VI

**MILESTONES IN THE DEVELOPMENT OF THE TREE-SEARCH MIMO DETECTORS: BEST-FIRST TYPE**

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>Fukatani et al. [321]</td>
<td>Applied Dijkstra’s algorithm [451] for reducing the complexity of the SD based MIMO detector at the expense of an increased storage size.</td>
</tr>
<tr>
<td>2006</td>
<td>Murugan et al. [66]</td>
<td>Proposed a unified framework for tree search decoding, which encompasses all existing SDs as special cases, hence unifying the depth-first search, the breadth-first search and the best-first search based on the proposed framework.</td>
</tr>
<tr>
<td>2012</td>
<td>Chang et al. [72]</td>
<td>A generalization of Dijkstra’s algorithm was developed as a unified tree-search detection framework; the proposed framework incorporates a parameter triplet that allows the configuration of the memory usage, detection complexity and the sorting dynamic associated with the tree-search algorithm; by tuning the different parameters, beneficial performance-complexity tradeoffs are attained and a fixed-complexity version can be conceived.</td>
</tr>
<tr>
<td>2012</td>
<td>Shen et al. [440]</td>
<td>Proposed the algorithms and VLSI architectures for both the best-first soft- and hard-decision tree-search based MIMO decoders in the context of a $4 \times 4$ 64-QAM system using 65-nm CMOS technology at 333 MHz clock frequency.</td>
</tr>
</tbody>
</table>
TABLE VII
MILESTONES IN THE DEVELOPMENT OF THE LR-BASED MIMO DETECTORS

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1982</td>
<td>Lenstra et al. [460]</td>
<td>First proposed the LLL algorithm for LR, which becomes the most popular LR algorithm used in practice.</td>
</tr>
<tr>
<td>2002</td>
<td>Yao et al. [83]</td>
<td>First applied the LR technique in conjunction with the traditional linear ZF and nonlinear ZF-SIC detector, showing a substantial performance gain at a modest additional computational complexity.</td>
</tr>
<tr>
<td>2003-2004</td>
<td>Windpassinger and Wubben et al. [84], [85]</td>
<td>Presented LR-aided MIMO detectors relying on real-valued LLL algorithms.</td>
</tr>
<tr>
<td>2003-2004</td>
<td>Windpassinger et al. [84], [86]</td>
<td>Proposed a real-valued LLL-based LR algorithm, which enables the application of the algorithm in MIMO systems having arbitrary numbers of dimensions. It was also shown that LR can be favorably applied in MIMO systems that use precoding.</td>
</tr>
<tr>
<td>2004</td>
<td>Wubben et al. [85], [466]</td>
<td>Extended the LR-aided linear ZF and nonlinear ZF-SIC MIMO detectors to their MMSE based counterparts.</td>
</tr>
<tr>
<td>2007</td>
<td>Taherzadeh et al. [87]</td>
<td>Demonstrated that the LLL based LR algorithm is capable of achieving full receive diversity of MIMO decoding.</td>
</tr>
<tr>
<td>2007-2008</td>
<td>Ling and Jaldén et al. [88], [89]</td>
<td>Provided upper bounds for the average computational complexity of the LLL algorithm, namely $O(N_1^3 \log N_t)$ and $O(N_2^2 \log \frac{N_r}{N_t-N_t+1})$, respectively.</td>
</tr>
<tr>
<td>2007-2008</td>
<td>Seethaler and Zhang et al. [90], [91]</td>
<td>Studied the performance of the Seyssen’s algorithm based LR techniques in MIMO detection problems.</td>
</tr>
<tr>
<td>2007</td>
<td>Burg et al. [92]</td>
<td>The first VLSI implementation of the LR technique relying on Brun’s algorithm was reported.</td>
</tr>
<tr>
<td>2008</td>
<td>Gestner et al. [94]</td>
<td>The first VLSI implementation of the LR technique relying on the complex-valued LLL algorithm was reported.</td>
</tr>
<tr>
<td>2005-2009</td>
<td>Guan and Ma et al. [95]–[98]</td>
<td>Proposed a number of complex-valued LLL algorithms which can be directly used in the complex-valued MIMO system model.</td>
</tr>
<tr>
<td>2006 - 2010</td>
<td>Silvola, Qi, Ponnampalam and Zhang et al. [99]–[102]</td>
<td>Studied a range of LR-aided soft-output MIMO detectors, including LR aided K-best [100], LR-aided MAP [99], LR-aided fixed radius algorithm, fixed candidates algorithm, fixed memory-usage algorithm etc. [102].</td>
</tr>
<tr>
<td>2013</td>
<td>Zhou et al. [103]</td>
<td>Proposed a class of element-based LR algorithms, which reduce the diagonal elements of the noise covariance matrix of linear detectors and thus enhance the asymptotic performance of linear detectors, in large-scale MIMO systems having hundreds of BS antennas.</td>
</tr>
</tbody>
</table>

target tracking and surveillance in a cluttered environment, where measurements are unlabelled and may be spurious. To elaborate a little further, it was developed for solving the problem of plot-to-track association in a radar tracker. In this context, all of the potential candidates for association to a specific track are combined into a single statistically most likely update, taking account of the statistical distributions of both the tracking errors and the clutter, while assuming that only one of the candidates is the desired target with the rest of them representing false alarms. A major extension of the JPD filter is the joint probabilistic data association (JPDA) filter [468], [469], which takes account of the situation that multiple targets are present out of all the potential candidates, and hence seeks to compute the joint decision probabilities for the multiple targets. In addition to their wide applications in radar, sonar, electro-optical sensor networks and navigation systems [467]–[481], the PDA techniques have also been applied in the field of computer vision for solving the visual target tracking problem [482]–[485].

The PDA approach may also be applied for solving challenging problems in digital communications. For example, it may be developed as a reduced-complexity design alternative to the optimal soft-decision aided MAP detectors/equalizers of MIMO channels [104]–[133], and it is also applicable to channel estimation of MIMO systems [486], [487]. Since we mainly focus on MIMO detection in this paper, a more detailed discussion of the PDA-based MIMO channel estimation will not be included in the sequel. As far as the PDA-based MIMO detection is concerned, it is Luo et al. [104] who first applied the PDA approach to the MUD problem of BPSK-modulated synchronous CDMA systems in 2001, showing a near-optimum performance at a significantly lower computational complexity than the ML detector. Thereafter, the PDA-based detector was naturally extended to the scenario of BPSK-modulated asynchronous CDMA systems [106], [107]. Recently, it was also extended to the symbol detection of QAM-aided SDM-MIMO systems [111]–[113], [116], [120], [130], striking an attractive tradeoff between the attainable performance and the complexity imposed. More specifically, in [111] a real-valued PDA (RPDA) was formulated for $M$-QAM constellations, which is based on the equivalent real-valued MIMO signal model previously discussed in Section VII. Additionally, in [112] an approximate complex-valued PDA (A-CPDA) detector was proposed, in which the complex-
valued Gaussian distribution is approximately characterized by a matched mean and a matched covariance only. Furthermore, the pseudo-covariance, as defined by Neeser and Massey in [301], was employed in [120] to fully characterize the complex-valued Gaussian distribution, and the resultant formulation of complex-valued PDA (CPDA) [120] was shown to outperform both the RPDA [111] and the A-CPDA [112].

In these PDA-based MIMO detectors/equalizers, the probabilities of the potential candidate symbols serve as the soft input/output information and are typically estimated relying on a self-iterative process. The key operation in this process is the iterative approximation of the interference-plus-noise term obeying a multimodal Gaussian mixture distribution by an ever-updated multivariate Gaussian distribution [104], [111]–[113], [129]. Therefore, the performance of the PDA based MIMO detectors is largely determined by the accuracy of the iterative Gaussian approximation, whose impact on the performance of the PDA based detectors was investigated in [116]. In order to further improve the accuracy of the Gaussian approximation, the authors of [127] proposed a PDA detector for correlated source bits using the joint detection of multiple consecutive symbol vectors. Additionally, in [130], [488] a unified direct bit-based PDA approach was proposed for detecting linear mapping based high-order rectangular QAM symbols, which achieves a better performance at a lower computational complexity than the CPDA detector of [120]. Furthermore, the PDA based iterative receiver design of FEC-coded MIMO systems was investigated in [132], [133], [489], [490], where it was revealed that the outputs of the conventional PDA detectors in [104], [111], [112], [120], [127], [130] are indeed the normalized symbol likelihoods, rather than the true APPs. Based on this insight, a pair of PDA based MIMO iterative receivers, namely the approximate and the exact Bayesian theorem based iterative PDA receivers were proposed in [132], [489] and [133], [490], respectively. Additionally, a distributed soft combining based PDA receiver was conceived in [131], [218] for BS cooperation aided multi-cell multiuser MIMO systems.

The advantages of the PDA based detectors are summarized as follows.

- First, it may achieve a near-optimal detection performance in certain circumstances, for example in the context of FEC-uncoded CDMA systems [104]–[107].
- Second, it has a low complexity that increases no faster than $O \left( M_i N_i^4 \right)$ per symbol vector, where $M_i$ is the number of PDA iterations, while $N_i$ represents either the number of users in CDMA [104]–[107], or the number of transmit antennas in multi-antenna aided MIMO systems [111], [112], [120].
- Third, it is inherently an soft-input soft-output algorithm, which is eminently applicable in combination with FEC codes such as convolutional codes, turbo codes [491], [492] or low-density parity-check (LDPC) codes [493], [494].
- Furthermore, the higher the number of transmit antennas or users, the better its performance, provided that the
channel is not overloaded ($N_I > N_O$) or rank-deficient [116]. However, due to its nature of approximation and iteration, the PDA based MIMO detector has not been well-understood compared to other mature MIMO detectors.

For the sake of more explicitly clarifying the fundamental principle of the PDA based MIMO detector, its Gaussian approximation process is conceptually illustrated in Fig. 23, which is based on the assumption that the interference-plus-noise term to be processed by the PDA detector obeys a single-variate multimodal (four-modal) Gaussian mixture distribution of $p_M(x) = p_1 \times f_1(x) + p_2 \times f_2(x) + p_3 \times f_3(x) + p_4 \times f_4(x)$. Here, the “single variate” assumption indicates that only a single interfering symbol, say $s_i$, exists for the other symbol to be detected. In other words, a $(2 \times 2)$-element VBLAST system is assumed. More specifically, the four-modal distribution observed in Fig. 23 stands for the case of a 4PAM-like scenario, which is a simplified real-valued example for $M$-QAM. More specifically, $p_M(x)$ is constructed by a mixture of four constituent Gaussian distributions $f_1(x)$, $f_2(x)$, $f_3(x)$, $f_4(x)$ having the same variance, but different means of $m_1 = -3$, $m_2 = -1$, $m_3 = 1$, $m_4 = +3$ and different constituent probabilities of $p_1$, $p_2$, $p_3$, $p_4$. The four constituent probabilities correspond to the different probabilities that an interfering symbol has the value $s_i = -3$, $s_i = -1$, $s_i = 1$ and $s_i = +3$, respectively.

The main contributions to the development of the PDA based MIMO detectors are summarized in Table VIII.

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>Luo et al. [104]</td>
<td>First applied the PDA filter technique to the MUD problem of synchronous CDMA systems, showing a near-optimum performance at a significantly reduced complexity.</td>
</tr>
<tr>
<td>2002</td>
<td>Pham et al. [106]</td>
<td>Proposed a PDA-Kalman MUD approach for asynchronous CDMA systems.</td>
</tr>
<tr>
<td>2003</td>
<td>Luo et al. [107]</td>
<td>Conceived a sliding-window PDA based MUD approach for asynchronous CDMA systems.</td>
</tr>
<tr>
<td>2003</td>
<td>Tan et al. [108]</td>
<td>Designed a PDA-based IDD receiver for a coded CDMA system using BPSK modulation.</td>
</tr>
<tr>
<td>2004</td>
<td>Pham et al. [111]</td>
<td>Extended the PDA detector to SDM-MIMO systems based on a real-valued signal model.</td>
</tr>
<tr>
<td>2004</td>
<td>Liu et al. [112]</td>
<td>Proposed a PDA-based soft equalization scheme for frequency-selective MIMO channels.</td>
</tr>
<tr>
<td>2005</td>
<td>Liu et al. [113]</td>
<td>Extended the PDA-Kalman MUD approach of [106] to the soft equalization of frequency-selective MIMO channels.</td>
</tr>
<tr>
<td>2005</td>
<td>Latsoudas et al. [114]</td>
<td>Proposed a hybrid MIMO detector that combined the SD and the PDA detectors.</td>
</tr>
<tr>
<td>2005</td>
<td>Fricke et al. [116]</td>
<td>Studied the impact of Gaussian approximation on the performance of the PDA based MIMO detector.</td>
</tr>
<tr>
<td>2006</td>
<td>Jia et al. [120]</td>
<td>Proposed a complex-valued PDA (CPDA) detector which takes the pseudo-covariance into account during the derivation of the complex-valued PDA detector.</td>
</tr>
<tr>
<td>2008</td>
<td>Kim et al. [126]</td>
<td>Applied the PDA method as a component of an iterative receiver designed for non-coherent MIMO systems.</td>
</tr>
<tr>
<td>2008</td>
<td>Yang et al. [127]</td>
<td>Proposed a PDA detector for correlated source bits using joint detection of multiple consecutive symbol vectors.</td>
</tr>
<tr>
<td>2009</td>
<td>Mohammed et al. [128]</td>
<td>Proposed a PDA detector for correlated source bits using joint detection of multiple consecutive symbol vectors.</td>
</tr>
<tr>
<td>2011</td>
<td>Yang et al. [130]</td>
<td>Proposed a unified direct bit-based PDA approach for detecting linear mapping based high-order rectangular QAM symbols, achieving a better performance at a lower computational complexity than the CPDA detector of [120].</td>
</tr>
<tr>
<td>2011</td>
<td>Yang et al. [131]</td>
<td>Proposed a distributed soft combining based PDA receiver for BS cooperation aided multi-cell multiuser MIMO systems.</td>
</tr>
<tr>
<td>2013</td>
<td>Yang et al. [132], [133]</td>
<td>Investigated the PDA based iterative receiver design for FEC-coded MIMO systems; revealed that the outputs of the conventional PDA detectors are indeed the normalized symbol likelihoods rather than the true APPs; proposed a pair of PDA based MIMO iterative receivers, namely the approximate and the exact Bayesian theorem based iterative PDA receivers.</td>
</tr>
</tbody>
</table>
G. Semidefinite Programming Relaxation Based Detectors

In contrast to other MIMO detectors, the SDPR approach relies on a relaxation of the optimum MIMO detection problem to the mathematical model of semidefinite programming (SDP) [323], [324], [495], which is a subfield of convex optimization theory [496].

Convex optimization constitutes a subfield of the generic mathematical optimization problem. It studies the minimization of a convex objective function over convex sets. Fig. 24 illustrates the basic framework of solving mathematical optimization problems using convex optimization. If a mathematical optimization problem is identified as a convex optimization problem, it is mathematically regarded as an easy problem, because powerful numerical algorithms, such as the interior-point methods [497], exist for efficiently finding the optimal solution of convex problems. Therefore, in mathematical optimization theory, the dividing line between the family of easy and difficult problems is convex versus nonconvex, rather than linear versus nonlinear. In other words, convex optimization problems are efficiently solvable, whereas nonconvex optimization problems are generally difficult to solve. Convex optimization has a range of other important properties. For example, in convex optimization problems, every locally optimal solution constitutes the globally optimal solution, hence there is no risk of being trapped in a local optimum. Additionally, a rigorous optimality condition and a duality theory exist for verifying the optimal nature of a solution in convex optimization problems. For more details of convex optimization, please refer to [324], [496].

The SDPR based MIMO detectors have recently received substantial research attention [134]–[146]. The most attractive characteristic of the SDPR-aided detectors is that they guarantee a so-called polynomial-time\(^{16}\) worst-case computational complexity, while achieving a high performance in certain circumstances. Most of the existing SDPR detectors are dependent on the specific modulation constellation. To elaborate a little further, SDPR was first proposed for a BPSK-modulated CDMA system [134], [135], [498]–[501], and then it was extended to quadrature phase-shift keying (QPSK) [136]. Simulation results showed that the SDPR detector is capable of achieving a near-ML BER performance, when using BPSK [134] and QPSK [136]. The numerical and analytical results of [137], [138] confirmed that the SDPR detector achieves the maximum possible diversity order, when using BPSK for transmission over a real-valued fading MIMO channel. The SDPR approach was also further developed for high-order modulation schemes, such as for \(M\)-ary PSK scenario in [139], [140], and for high-order rectangular QAM in [141]–[145]. As for the high-order QAM scenario, it was recently shown in [146] that the so-called polynomial-inspired SDPR (PI-SDPR) [141], the bound-constrained SDPR (BC-SDPR) [143] and the virtually antipodal SDPR (VA-SDPR) [145] are actually equivalent in the sense that they arrive at the same symbol decisions, and hence they exhibit an identical SER performance.\(^{17}\) Furthermore, a bit-based SDPR detector capable of directly detecting the nonlinear Gray mapping aided rectangular high-order QAM symbols was proposed in [147], where the unequal error protection property (UEP) of QAM bits was exploited and the resultant SDPR detector was shown to outperform that of [145]. It should be noted, however, that for high-order modulation scenarios, the performance of the SDPR detectors is not as promising as that of the BPSK/QPSK scenario. Therefore, there is a need to further improve the performance of the SDPR based MIMO detector designed for high-order QAM constellations, while maintaining its appealingly low computational complexity. The basic principle of SDPR based detectors is illustrated in Fig. 25, where the blue boxes represent the technical challenges. Furthermore, the main contributions to the development of the SDPR based MIMO detectors are summarized in Table IX.

H. Detection in Rank-Deficient and Overloaded MIMO Systems

For MIMO detection, typically it is preferable to have a full-rank channel matrix, namely \(\text{rank}(H) = N_t\) or \(N_o\). In CDMA systems, this requirement may be satisfied by using well-designed spreading codes. In multi-antenna SDM systems, when an ideal rich scattering multipath environment is assumed, typically independently fading communications channels are encountered between each transmit/receive antenna pair. Then, the full-rank requirement may also be satisfied. However, in some propagation scenarios, the channel matrix \(H\) may not be of full-rank. For example, if the spatial

\(^{16}\)The computational complexity increases as a polynomial function of \(N_t\).

\(^{17}\)More specifically, the solution equivalence of the PI-SDPR and BC-SDPR schemes holds for 16-QAM and 64-QAM, while that between the BC-SDPR and VA-SDPR techniques holds for any \(4^q\)-QAM scheme, where \(q\) is a positive integer. The SDPR QAM detector of [144] exhibits a better performance than that of [141], [143], [145], but has a much higher complexity.
### Table IX

**Milestones in the Development of the SDPR-based MIMO Detectors**

<table>
<thead>
<tr>
<th>Year</th>
<th>Authors</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001 -2003</td>
<td>Tan et al. [134], Ma et al. [135], [498], [499], and Wang et al. [500], [501]</td>
<td>These authors independently proposed a SDPR based MUD for BPSK-modulated synchronous CDMA systems; the eigen-decomposition based method of [134], [498], [500], [501] and the randomization method of [135], [499] were proposed for converting the continuous-valued solution of the SDP problem into the binary decision output. Additionally, a cutting plane method was introduced for further improving the performance of the SDPR detector for systems supporting a large number of users [134], [498]; it was shown that the classic MUDs, such as the linear ZF/MMSE detector, can be interpreted as degenerate forms of the SDPR based MUD [135], [499].</td>
</tr>
<tr>
<td>2003</td>
<td>Steingrimsson et al. [502]</td>
<td>Proposed a soft SDPR detector for an IDD receiver of QPSK-aided MIMO systems employing LDPC codes.</td>
</tr>
<tr>
<td>2004</td>
<td>Ma et al. [136]</td>
<td>Conceived a SDPR based MUD for BPSK/QPSK aided asynchronous CDMA systems with multiple receive antennas in frequency-selective fading environments; based on a flexible block alternating likelihood maximization (BALM) principle, the large-scale ML detection problem was decomposed into smaller subproblems, and each subproblem was solved by the SDPR detector.</td>
</tr>
<tr>
<td>2003 -2004</td>
<td>Luo et al. [139] and Ma et al. [140]</td>
<td>Proposed SDPR detectors for general $M$-PSK aided synchronous CDMA systems.</td>
</tr>
<tr>
<td>2005</td>
<td>Kisialiou et al. [137]</td>
<td>Provided the first analytical study of the SDPR detector for BPSK-aided MIMO systems; it was shown that the SDPR detector is capable of achieving the same BER performance as that of the ML detector in high-SNR scenarios, while at the low SNR region, the SDPR detector serves as a constant factor approximation to the ML detector in large systems.</td>
</tr>
<tr>
<td>2005</td>
<td>Wiesel et al. [141]</td>
<td>Designed a PI-SDPR detector for 16-QAM aided MIMO systems, which can be extended to high-order $M$-QAM scenarios.</td>
</tr>
<tr>
<td>2006</td>
<td>Sidiropoulos et al. [143]</td>
<td>Advocated a BC-SDPR detector for employment in high-order $M$-QAM aided MIMO systems.</td>
</tr>
<tr>
<td>2007</td>
<td>Mao et al. [145]</td>
<td>Proposed a VA-SDPR detector for $M$-QAM aided multicarrier CDMA(MC-CDMA) systems; the method can directly operate at the bit-level in the context of linear mapping based $M$-QAM.</td>
</tr>
<tr>
<td>2007</td>
<td>Mobasher et al. [144]</td>
<td>Studied several variants of the SDPR detectors, and showed that it is possible to further improve the SDPR detector’s performance by increasing their complexity.</td>
</tr>
<tr>
<td>2008</td>
<td>Jaldén et al. [138]</td>
<td>Analytically demonstrated that the SDPR based detector is capable of achieving full receive diversity order in BPSK-aided real-valued MIMO channels.</td>
</tr>
<tr>
<td>2009</td>
<td>Ma et al. [146]</td>
<td>Demonstrated that the PI-SDPR of [141], the BC-SDPR of [143], and the VA-SDPR of [145] are actually equivalent in the sense that they obtain the same symbol decisions, and hence exhibit an identical SER performance.</td>
</tr>
<tr>
<td>2013</td>
<td>Yang et al. [147]</td>
<td>Proposed a bit-based SDPR detector capable of directly detecting the nonlinear Gray mapping aided rectangular high-order QAM symbols, where the unequal error protection property (UEP) of QAM bits was exploited and the resultant SDPR detector outperforms that of [145].</td>
</tr>
</tbody>
</table>

Separation between the antenna elements of the transmitter and/or the receiver is not large enough and hence the angular spread is small, the strong correlation between the antenna elements results in a rank-deficient channel matrix, i.e. we have $\text{rank}(\mathbf{H}) < \min(N_t, N_r)$. Hence, the spatial degrees of freedom available are reduced, which translates into a decreased MIMO capacity [503]–[508]. Furthermore, even if the spatial separation between antenna elements is sufficiently large, it is still possible that $\mathbf{H}$ is rank-deficient. This is due to the so-called “keyhole/pinhole effects” [508]–[513], which may be simply understood as a diffraction phenomenon, where a large obstacle with a small keyhole punched through it is placed between the MIMO transmitter and receiver, hence the only channel the radio wave can propagate through to the receiver is the keyhole. Due to this effect, the channel matrix $\mathbf{H}$ is degenerate and has only a single degree of freedom, i.e. we have $\text{rank}(\mathbf{H}) = 1$, even though the entries of $\mathbf{H}$ are uncorrelated.

Another preferable condition for the detection in CDMA and SDM-MIMO systems is that the system is not overloaded.
Then, the channel matrix is “fat” and does not have full column-rank (but it may still have full row-rank.). In the multi-antenna scenario, this means that $N_t \leq N_r$, while in CDMA systems, it means that the number of users is higher than the dimension of the signal space/the processing gain of the system. As far as MIMO detection is considered, both the rank-deficient scenario discussed above and the overloaded scenario face the common challenge that the standard versions of most of the representative MIMO detectors, such as the linear ZF/MMSE detector, the ZF/MMSE-SIC detector and the original linear-decorrelation based PDA detector [104], [111] that invoke the inverse of $\mathbf{H}$, and the standard SD detector that invokes standard QRD or Cholesky’s factorization [53], usually provide an unacceptably poor performance, because they are invoked for finding the solution of an under-determined linear system subject to random noise.

Several strategies have been proposed to circumvent this predicament, such as the “pseudo-inverse” based linear detection [180], [239], the group detection [514]–[519], the generalized SD detector [56], [57], [520]–[527], the modified non-decorrelated PDA detection [112], [116], [121], [130], the modified SDPR detection [147], [526], [528], the metaheuristics based detection [529]–[550] etc. It is possible to design various scenario-dependent MIMO detectors for the rank-deficient and overloaded MIMO systems. However, it seems that the group detection strategy [514]–[519] and the search-based detection, regardless of the ML detector, the generalized SD detector [56], [57], [520]–[527] and the metaheuristics based detector [529]–[550], are particularly suitable for rank-deficient and overloaded MIMO scenarios.

I. Impact of Soft-Decision and Transmit Preprocessing on MIMO Detection

In previous sections we aim for understanding the fundamental properties of MIMO detection algorithms. However, if we look at the entire process of communication, the assumption that only the receiver is responsible for signal recovery represents a passive and incomplete strategy. In fact, almost all practical systems invoke some form of encoding or transmit preprocessing, such as FEC, space-time coding and precoding/beamforming, to actively improve the performance of signal recovery or to reduce the detector’s computational complexity from the transmitter side.

To elaborate a little further, when FEC is used, tremendous efforts have been devoted to designing soft-input soft-output MIMO detectors that can fit into the powerful “turbo processing principle” [491], [492], [551]–[553] based IDD receiver architecture conceived for achieving near-optimum performance. All the MIMO detectors reviewed in Section VIII have their soft-decision versions to fit into IDD receivers. The representative contributions to iterative MIMO detection and decoding include: the optimal MAP detector based iterative receiver [554]–[558], the expectation-maximization (EM) algorithm based soft-decision MUD [559], the soft-decision MMSE detector assisted iterative receiver [560]–[562], the soft-decision SD based MIMO iterative receiver [65], [73], [74], [81], [525], [563], the PDA detector based iterative receiver [108], [125], [132], [133], the soft-decision SDPR detector aided MIMO iterative receiver [502], [564], the soft-decision multiple symbol differential SD (MSDSD) detector based non-coherent iterative receiver [565], and the soft-decision iterative receiver for LS-MIMO systems [339], [440] discussed in Section IX. Yet another important contribution to IDD design is the extrinsic information transfer (EXIT) chart invented by ten Brink [566], [567], which is a powerful tool conceived not only for analyzing the convergence behavior of iterative receivers, but also for assisting near-capacity wireless system design [74], [77], [342], [568]–[576]. For more details on designing iterative MIMO receivers, please refer to [74], [342], [555], [557], [560], [577]–[580].

Additionally, when space-time coding is employed, as we discussed in Section IV, the optimal ML decoding can be implemented with a simple separate symbol-by-symbol decoding strategy for orthogonal STBCs and with a linear-complexity pairwise decoding strategy for quasi-orthogonal STBCs [274]. As a result, the MIMO detection problem does not constitute a grave challenge for STBC aided MIMO systems. Similarly, when precoding/beamforming techniques [84], [581]–[587] are employed in SDM-MIMO systems, the interference between the transmit antennas may be significantly mitigated or even completely removed (when using ZF-based linear precoding). As a result, the signal detection task of a precoded MIMO system becomes less challenging compared to that of SDM-MIMO systems invoking no preprocessing. The key insight gained here is that we can design an encoder or precoder to improve the performance or to reduce the computational complexity of decoders/detectors.

J. Guidelines on Choosing the Right MIMO Detectors

As we mentioned in Section VIII-A2, the optimality of MIMO detectors strongly depends both on the criteria of “goodness” and on the assumptions made for specific application scenarios. Each type of MIMO detector has a different performance-and-complexity profile18, and each of them has its own pros and cons. Therefore, in general there is no simple answer as to which algorithm is the best. In what follows, we first provide a qualitative comparison of the performance and complexity characteristics of the MIMO detectors reviewed, and then summarize their analytical performance and complexity results in Table X.

- The MAP/ML based MIMO detectors relying on brute-force search have the optimal VER performance (not necessarily optimal BER or SER) and a computational complexity which increases exponentially with the system’s dimension (e.g. the number of transmit antennas or users). Naturally, their computational complexity order $O(M^{N_t})$ is the highest amongst all the MIMO detectors. Additionally, the MAP/ML algorithms have to be aware of the amplitudes of the transmitted symbols for

18Generally, “performance” and “complexity” may be interpreted in various ways. For example, the “performance” can be error probability, robustness to system imperfections, configuration flexibility, application generality etc., while the “complexity” could be computational complexity, hardware/silicon complexity and so on.
### TABLE X

**Performance and complexity comparison of various hard-decision MIMO detectors in uncoded SDM-MIMO systems, where \( N_r \geq N_t \).**

<table>
<thead>
<tr>
<th>Detector</th>
<th>Receive diversity order at high SNR</th>
<th>Error probability/DMT/asymptotic analysis</th>
<th>SNR penalty</th>
<th>Worst-case computational complexity order per symbol vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP/ML</td>
<td>( N_r ) [356], [358]</td>
<td>See [237], [354], [356]-[360]</td>
<td>Zero</td>
<td>( O(M^{N_t}) )</td>
</tr>
<tr>
<td>Linear ZF</td>
<td>( N_r - N_t + 1 ) [588]</td>
<td>See [381], [384], [385], [588]</td>
<td>High</td>
<td>Between ( O(N_t^2) ) and ( O(N_t^3) )</td>
</tr>
<tr>
<td>Linear MMSE</td>
<td>( N_r - N_t + 1 ) [362], [363], [588]</td>
<td>See [378], [588], [362], [363], [381]-[384]</td>
<td>Lower than that of linear ZF</td>
<td>Between ( O(N_t^2) ) and ( O(N_t^3) )</td>
</tr>
<tr>
<td>ZF/MMSE-SIC</td>
<td>( N_r - N_t + 1 ) [44], [385], [589]-[591]</td>
<td>See [44], [385], [589]-[591]</td>
<td>Lower than that of linear ZF/MMSE</td>
<td>Between ( O(N_t^2) ) and ( O(N_t^3) )</td>
</tr>
<tr>
<td>ZF/MMSE-OSIC</td>
<td>( N_r - N_t + 1 ) [385], [589], [591], [592]</td>
<td>See [385], [589], [591], [592]</td>
<td>Lower than that of ZF/MMSE-SIC</td>
<td>Between ( O(N_t^2) ) and ( O(N_t^3) )</td>
</tr>
<tr>
<td>SD</td>
<td>( N_r ) [593]</td>
<td>The same as that of ML, if it is used for obtaining the exact ML solution</td>
<td>(Can be) zero in general</td>
<td>( O(M^{\beta N_t}) ), ( \beta \in (0, 1] ) [58]-[60], [594], [595]</td>
</tr>
<tr>
<td>FCSD</td>
<td>( \min(N_r, (N_r - N_t)(p + 1) + (p + 1)^2) ) where the first ( p ) levels experience full search [82], [596]</td>
<td>See [82], [596]</td>
<td>Approach zero at high-SNR</td>
<td>( O(M\sqrt{N_r}) ) [82], [596]</td>
</tr>
<tr>
<td>K-best SD</td>
<td>( N_r - N_t + 1 ) to ( N_r ), depending on the value of ( K )</td>
<td>unknown for arbitrary ( K ), flexible and suitable for VLSI implementation</td>
<td>Between that of SIC and ML, depending on ( K ) [63], [449]</td>
<td>Between that of SIC and ML, depending on ( K )</td>
</tr>
<tr>
<td>LLL-LR-ZF/MMSE/SIC</td>
<td>( N_r ) [87], [96]-[98], [597]</td>
<td>See [87], [96]-[98], [597]-[601]</td>
<td>Can approach zero, but the actual gap depends on how well the particular channel can be reduced.</td>
<td>( O(M_i N_t^4) ) to ( O(M_i N_t^4) ) where ( M_i ) is the number of PDA iterations</td>
</tr>
<tr>
<td>PDA</td>
<td>unknown</td>
<td>See [116], [118], [249], [121], [132], [602]</td>
<td>Approach zero for large ( N_t ) [104], [116], [118]</td>
<td>( O(M_i N_t^4) ) to ( O(M_i N_t^4) ) where ( M_i ) is the number of PDA iterations</td>
</tr>
<tr>
<td>SDPR</td>
<td>( N_r ) only for BPSK transmission over real-valued Gaussian fading channels [138]</td>
<td>See [135], [137], [603], [138], [146], [604]</td>
<td>Typically near-ML for BPSK/QPSK, but not for high-order constellations unless complexity is substantially increased [144]</td>
<td>Constellation-dependent: ( O((1 + N_t \log_2 M)^{3.5}) ) [140], [502], [605], [143], [147], [323], [606] to ( O((\sqrt{M}(2N_t + 1))^{6.5}) ) [143], [144], [146]</td>
</tr>
</tbody>
</table>
calculating the decision statistics. However, the MAP/ML detector is insensitive to channel imperfections and rank-deficiency/overloading, and it has the best possible error probability performance across the entire SNR region. When the system’s dimension is not too large, it remains possible to implement the exact MAP/ML algorithm in practical systems with the aid of state-of-the-art VLSI technologies.

- The linear MIMO detectors typically have the lowest computational complexity between $O(N_t^2)$ and $O(N_t^3)$, although there exist subtle differences amongst the computational complexities of different linear detectors. Naturally, in general they have the least attractive error probability performance. However, in some scenarios, such as the large-scale MIMO systems to be detailed in Section IX, where the receiver side has a significantly higher number of antennas than the transmitter side, the linear MF, ZF, MMSE, MBER etc. based MIMO detectors may achieve a near-ML error probability. Additionally, the linear MF and ZF detectors only have to know the channel matrix $H$, but the linear MMSE detector additionally has to estimate the noise variance. Furthermore, as indicated in Section VIII-A2, the linear ZF detector is preferable in interference-dominated scenarios, the linear MF detector is preferable in noise-dominated scenarios, while the linear MMSE detector provides the highest SINR amongst all linear detectors, which makes it preferable in scenarios where the noise and the interference have a comparable level. Finally, the linear ZF and MMSE detectors exhibit an inadequate performance in rank-deficient/overloaded systems, where the number of independent inputs is higher than the dimension of the received signals, while the linear MF detector remains applicable.

- The interference cancellation based MIMO detectors have a computational complexity between $O(N_t^3)$ and $O(N_t^4)$, and typically they have a much more attractive error probability performance than the linear detectors. Theoretically, the SIC/DFD based detectors are capable of approaching the Shannon capacity, provided that there is no error propagation at any of the decision stages. By contrast, the PIC/MIC based detectors do not have this property. Compared to PIC/MIC, the SIC/DFD detectors are more sensitive to error propagation. However, this makes them preferable in the “near-far” scenario, where the powers of different users are significantly different, such as those of the cell-center user and cell-edge user. Furthermore, the SIC/DFD detectors may have a higher processing delay than the PIC detectors. Additionally, similar to the linear ZF and MMSE detectors, the interference cancellation based detectors are not generally applicable to the rank-deficient/overloaded scenarios.

- The tree-search based MIMO detectors, especially the $K$-best detectors, have the flexibility to achieve different error probability versus computational complexity trade-offs. They are even capable of attaining the optimum ML performance at a reduced complexity. In contrast to other types of MIMO detectors, the tree-search based detectors typically have a non-deterministic complexity, which is a challenge for hardware implementation, albeit it is possible to design fixed-complexity tree-search detectors. Therefore, the average computational complexity, worst-case computational complexity and even the computational complexity distribution become important complexity metrics to consider. Note, however, that theoretically the tree-search based MIMO detectors still have an exponentially increasing worst-case/average computational complexity, in which case the exponent depends on different system parameters, such as the noise variance. As a result, the tree-search based detectors may not be suitable for low-SNR scenarios. Additionally, it may be possible to design tree-search based detectors for rank-deficient/overloaded scenarios. Furthermore, the tree-search based detectors rely on specific enumeration strategies, which by nature are not suitable for large-scale MIMO systems that have a high number of inputs.

- The LR algorithms constitute a family of powerful preprocessing techniques conceived for improving the “quality” of the effective channel matrix. They can be used in conjunction with all the other MIMO detectors. Since practically usable LR algorithms, such as the LLL algorithm, have a polynomially increasing computational complexity, the LR based MIMO detectors do not have a significantly increased total computational complexity. Hence, LR may be particularly useful for designing high-performance MIMO detectors maintaining a low complexity, which is critical in numerous practical implementations. However, the LR techniques do not fundamentally change the pros and cons of their baseline detectors.

- Compared to the other MIMO detectors mentioned above, the SDPR and PDA based MIMO detectors are not well-understood at the time of writing and they have not achieved the same degree of practical success, which is partially indicated by the lack of VLSI implementations of these two types of detectors. Although SDPR detectors have a favorable worst-case polynomial complexity, which is roughly between $O(1 + N_t \log_2 M)^{3.5}$ and $O((\sqrt{N_t} + 1)^{6.5})$, their achievable error probability performance becomes less attractive for high-order modulations (but they may achieve near-ML performance for BPSK and QPSK constellations). The Gaussian-mixture approximation based PDA detectors operate in a way similar to the classic soft interference cancellation, hence their computational complexity is similar to that of the soft SIC detectors, i.e. typically on the order between $O(M_t N_t^2)$ and $O(M_t N_t^3)$. As a result, the PDA detectors are also sensitive to error propagation, whilst exhibiting the nice property of preferring a large number of inputs, provided that the receive dimensions are no less than that of the inputs. Hence, for certain large-scale MIMO scenarios, both SDPR and PDA based detectors may be attractive. Finally, for large-scale MIMO systems which have a similarly large number of transmit and receive antennas, it might be valuable to resort to metaheuristics based algorithms, since all the other MIMO detectors might either be excessively complex or fail to provide a high performance. Some of the metaheuristics based
large-scale MIMO detectors are described in Section IX-D.

**IX. DETECTION IN LS-MIMO SYSTEMS**

Having reviewed the representative families of MIMO detection algorithms in Section VIII, let us now shift our attention to the detection problem encountered in the emerging massive/LS-MIMO systems [25], [373], [607], where dozens or even hundreds of antennas may be invoked and an unprecedented spectral efficiency/diversity order may be achieved. The major benefits of LS-MIMOs can be deduced from the following well-known results. For transmission over a quasi-static channel where a codeword occupies only a single coherence-time and coherence-bandwidth interval, the outage probability of a point-to-point MIMO link scales according to

\[
\Pr_{\text{outage}} \propto \text{SNR}^{-N_t N_r},
\]

which indicates that potentially a diversity order of \((N_t \times N_r)\) may be achieved. In other words, the MIMO link’s reliability quantified in terms of its error rate falls exponentially with \(N_t\) and/or \(N_r\) when SNR increases. Additionally, on a fast-fading MIMO channel, the achievable rate scales as

\[
\min\{N_t, N_r\} \log_2(1 + \text{SNR}),
\]

which indicates that the achievable rate of a MIMO system scales linearly with \(\min\{N_t, N_r\}\), and hence it is possible to attain a high data rate using a large \(N_t\) and/or \(N_r\). In conclusion, fundamentally, using more antennas grants us higher degrees of freedom in the spatial domain without increasing the bandwidth occupied.

The LS-MIMO systems can be implemented in a variety of ways. For example, in the operational 3G/4G wireless communication systems, a point-to-point LS-MIMO system might be constructed to provide high-throughput wireless backhaul connectivity between the BSs by using a large number of antennas at each BS. However, apart from this particular application, it is typically quite challenging to construct a point-to-point LS-MIMO system where the antenna elements can have a sufficiently high spatial separation to guarantee a well-conditioned channel matrix. Furthermore, achieving the attractive multiplexing gains promised by point-to-point LS-MIMO schemes requires a high SNR. On the other hand, a multiuser LS-MIMO system [608], [609] can be envisaged, where the BS may be equipped with hundreds of antenna elements and serves dozens of MSs each having only a few antennas. Additionally, the LS-MIMO may be implemented in the extremely high frequency (EHF) band (i.e. at millimeter wave (MMW) frequencies ranging from 30 to 300 GHz and having wavelengths spanning from ten to one millimeter [33]). They may also be considered in the optical band for frequencies ranging from 300 GHz to 300 PHz and including the infrared, the visible and the ultraviolet band [610], [611]. Due to the adverse propagation properties, the coverage of the LS-MIMO systems operating in these high-frequency bands might be significantly limited, hence they are more applicable to indoor environments [610] or small-cell scenarios [612]. For the sake of more explicit clarity, several typical antenna configurations and deployment scenarios of LS-MIMO systems [613]. To elaborate a little further, the simplest linear array propagates signals on the two-dimensional plane and it typically occupies a large physical area. By contrast, the rectangular, cylindrical and spherical arrays are capable of radiating signals to any directions in the three-dimensional space. These antenna arrays are more complex, but also more compact, hence occupying a smaller physical area. Additionally, a virtual LS-MIMO may be constructed relying on distributed antenna arrays, which may be exploited to enhance the indoor coverage or outdoor cooperation [613].

As pointed out in Section VIII-A, the key motivation of studying the fundamental MIMO detection problem is that the computational complexity of the optimum ML/MAP MIMO detection increases exponentially with the problem size. Therefore, in principle the MIMO detection problem has intrinsically embedded the “large-scale” concept. In this regard, people may argue that the detection in LS-MIMO systems is not a novel problem, and consequently the detectors conceived for LS-MIMO systems might have no significant difference with respect to the existing MIMO detectors, except for the associated larger problem size. At first glance, this seems to be true. However, due to the limitations of practical applications, in the past large-scale MIMO systems have been regarded as being impractical and most of the research focused on small-scale MIMO systems. Nonetheless, in addition to their significant link reliability and throughput benefits, the LS-MIMO systems have been shown to enjoy some distinct advantages that are not available in small-scale MIMO systems. These benefits are mainly attributed to a range of relevant results in random matrix theory [614], [615], and might be capable of circumventing signal processing problems in LS-MIMO

In a single-cell/noncooperative multi-cell MIMO system the BS is not concerned about the CCI imposed by the transmissions of other cells. In this scenario, as pointed out in Section III, the detection problems encountered in both the point-to-point MIMOs (see Fig. 8) and the multiple-access MIMOs (see Fig. 9) can be characterized using the same received signal model of (1). From the antenna configuration point of view, there are two types of single-cell/noncooperative multi-cell LS-MIMO systems. As shown in Fig. 27, in the Type-I system, a large number of collocated antennas may be mounted on the receiver, and also a large number of collocated or distributed antennas are used at the transmitter. Mathematically, the antenna configuration of the Type-I system may be characterized by

$$\lim_{N_t, N_r \to \infty} \frac{N_t}{N_r} = c$$

with $c$ being a positive constant. (26) indicates that both $N_t$ and $N_r$ tend to infinity at the same rate. By comparison, in the Type-II system, only the receiver is equipped with a large number of collocated antennas, while the total number of active antennas at the transmit side is significantly smaller. Hence, the antenna configuration of this system may be characterized as

$$\lim_{N_r \to \infty} \frac{N_t}{N_r} = 0.$$  

For the Type-I system, it has been shown that the empirical distribution of the singular values of the random channel matrix $H$ converges to a deterministic limiting distribution\(^{19}\) for almost all realizations of $H$, which is a result of the Marčenko and Pastur law [616]. In other words, as $H$ becomes larger (in terms of both $N_t$ and $N_r$), its singular values become less sensitive to the actual distributions of the i.i.d. entries of $H$, and the channel becomes more and more deterministic. The Marčenko and Pastur law also shows that as the size of $H$ increases, the diagonal entries of $H^H H$ become increasingly larger in magnitude than the off-diagonal entries. This is the so-called “channel-hardening” behavior, which may be exploited for large-scale MIMO detection. To be more specific, the matrix inversion invoked by many MIMO detectors such as the ZF-aided detector, the MMSE-aided detector and the PDA-aided detector etc., may be conveniently approximated using the series expansion technique for large-dimensional random matrices [401]. Additionally, the channel-hardening phenomenon may allow low-complexity detection algorithms to achieve a good performance for large-scale MIMO systems [617].

The Type-II system essentially deals with the MIMO detection problem encountered on an underloaded uplink\(^{20}\), as shown in Fig. 28. On the one hand, since the number of BS antennas may be significantly higher than the total number of active MS antennas, a very unbalanced antenna configuration is encountered, which results in a high receive diversity order. In the extreme case shown by (27), the receive diversity gain obtained is so high that the impact of both the MUI and the noise diminishes. Additionally, the channel vectors associated with distinct MSs may become asymptotically orthogonal. Furthermore, another beneficial result of the Marčenko and Pastur law [616] is that very tall (with large $N_r$) and very wide (with large $N_t$) channel matrices $H$ are very well conditioned. Therefore, in the Type-II system, even the simplest MF detector is capable of achieving a near-optimum performance [373], [607]. Similarly, when considering the precoding based downlink of the single-cell/noncooperative multi-cell TDD system, it was also revealed that increasing the number of BS antennas is always beneficial, even when the SNR is low and the channel estimate is poor. Furthermore, when the number of BS antennas tends to infinity, the effects of both the small-scale fast fading and uncorrelated noise are mitigated. In other words, a large number of BS antennas, regardless

\(^{19}\)This limiting distribution is the so-called quarter circle law [44, Chapter 8.2].

\(^{20}\)On the downlink, large-scale MIMO precoding techniques may be employed, which facilitates the employment of simple receivers at each MS, because the precoder is capable of eliminating the IAI at the transmitter with the aid of accurate channel knowledge.
of whether the uplink or the downlink is considered, may be exploited to trade for relevant performance improvements, such as compensating for the low SNR and/or poor channel estimates [25], [373], [607], [618].

However, in the noncooperative multi-cell scenario, due to the so-called “pilot contamination” problem [21] [25], [607], the interference emanating from other cells does exist and becomes the major limiting factor of the achievable performance [25], [607]. Therefore, in order to further enhance the achievable performance, the BS cooperation based multi-cell joint processing philosophy has to be adopted [131], as detailed below.

**B. Detection in Cooperative Multi-cell Multiuser LS-MIMO Systems**

As pointed out in Section III, the multi-cell transmission scenario is characterized by the so-called “MIMO interference channels” of Fig. 11. Fundamentally, in order to cope with the interference, it may be desirable to transform the distributed model (such as the BSs of multiple cells) to a centralized model. This may be achieved by centralized/distributed BS cooperation [24], [131], where multiple BSs of adjacent cells may be connected via high-capacity optic fiber or microwave links, as shown in Fig. 29. As a result, effectively a physical/virtual super-BS is constructed to serve the cluster of collaborative cells, and this physical/virtual centralized model provides the performance upper bound of the original distributed system model. As far as detection is concerned, in principle most of the detection algorithms developed for the single-cell/noncooperative multi-cell scenarios may be adapted to the uplink of the cooperative multi-cell LS-MIMO system. The BS cooperation aided network MIMO detectors may be designed based on two distinct philosophies, namely using either interference cancellation [215] or data fusion [131]. However, the employment of BS cooperation might result in substantially increased backhaul traffic, which represents one of the major challenges facing the BS cooperation aided network MIMO.

**C. Applicability of Existing MIMO Detection Algorithms to LS-MIMO**

An inherent characteristic of LS-MIMO systems is their large dimension. Before investigating the applicability of existing MIMO detection algorithms in the LS-MIMO context, we have to identify which specific type of LS-MIMO systems is considered. On the one hand, in general most of the existing MIMO detectors would be applicable to a Type-II LS-MIMO system, where it is possible that low-complexity linear MIMO detectors might be capable of achieving near-optimum performance. Hence, the employment of more sophisticated MIMO detectors, such as the SD detector, may become unnecessary. On the other hand, some existing MIMO detection algorithms that have been specifically tailored for conventional small-/medium-scale MIMO systems might not be applicable to the Type-I LS-MIMO systems. To elaborate a little further, the family of tree-search based MIMO detectors, such as the popular SD detector that has a worst-case computational complexity increasing exponentially with the number of transmit antennas (see Section VIII-D for more details), will become less feasible in the Type-I LS-MIMO systems. Nonetheless, it might still be invoked in the Type-II LS-MIMO systems. By contrast, the PDA algorithm [130]–[133], which invokes the central limit theorem to perform stochastic interference modelling and imposes a polynomial-time worst-case computational complexity, will achieve an attractive performance versus complexity tradeoff in the Type-I LS-MIMO systems. Similarly, the convex optimization based SDPR detectors, which also exhibit a polynomial-time worst-case complexity as a function of the number of transmit antennas, might potentially be applicable to the Type-I LS-MIMO systems [147].

**D. Recent Advances in LS-MIMO Detection**

The LS-MIMO systems have become a hot research topic following Marzetta’s seminal work [607]. However, in terms of detection, several earlier works had touched upon this topic from either a large system analysis or an asymptotic performance analysis perspective. To elaborate a little further, in 2006 Tan and Rasmussen [118] derived a class of asymptotically optimal nonlinear MMSE MUDs based on a multivariate Gaussian approximation of the MUI for large-scale CDMA systems. This approach provided an alternative analytical justification for the structure of the PDA based detectors. The associated performance analysis showed that the BER performance of the PDA detectors can be accurately predicted and is close to the optimal detector’s performance.

---

21This is essentially the interference caused by reusing pilot sequences in adjacent cells.

22Note that it is quite common to solve hundreds of unknown variables in a convex optimization problem.
for large CDMA systems. Also in 2006, Liang et al. [325] proposed a block-iterative generalized decision feedback equalizer (BI-GDFE) for LS-MIMO systems using PSK constellations. Their asymptotic performance analysis demonstrated that the BI-GDFE closely approaches the single-user matched-filter bound (MFB) for large random MIMO channels, provided that the received SNR is sufficiently high [326]. Furthermore, in 2007 Liang et al. derived both the limit and the asymptotic distribution of the SINR for a class of MMSE receivers invoked in large-scale CDMA systems supporting unequal-power users. Their solution relied on the random matrix theory. They also proved that the limiting SINR converges to a deterministic value when \( K \) is the number of transmit antennas, \( N \) the number of receive antennas, and \( c \) a positive constant. Recall that this insight is the same as that discussed in the context of (26).

Additionally, they proved that the SINR of each particular user is asymptotically Gaussian for large \( N \) and derived the closed-form expressions of the variance for the SINR variable under both real-valued spreading and complex-valued spreading.

Chockalingam et al. also made significant contributions to the LS-MIMO detection problem, mainly using a variety of metaheuristics based local search algorithms invoked from machine learning/artificial intelligence [330]. More specifically, in 2008 they extended the low-complexity likelihood ascent search (LAS) based MUD [619]–[621] to the Type-I LS-MIMO system having up to 600 transmit and receive antennas [176]. This detector relies on the local neighborhood search and has its roots in the family of Hopfield neural network (HNN) based MUD algorithms [624]–[629]. It was shown that the LAS detector is capable of achieving near single-input single-output AWGN performance in a fading LS-MIMO environment at an average per-bit complexity of \( O(N_t N_r) \) [176]. Subsequently, they applied another local neighborhood search algorithm, namely the reactive tabu search (RTS) algorithm, to the detection of LS-MIMO systems. The RTS detector was shown to perform better than the LAS detector, because it relied on an efficient local minima exit strategy [128]. Additionally, a class of belief propagation (BP) LS-MIMO detectors relying on graphical models were proposed in [331], [332], [337]. A range of other detectors were studied by Chockalingam and his team in the context of LS-MIMO systems, including the randomized Markov chain Monte Carlo (R-MCMC) detector [333], the randomized search (RS) detector [333], the Monte-Carlo-Sampling based detector which jointly relies on a mixed Gibbs sampling (MGS) strategy combined with a multiple restart (MR) strategy [334], and the LR based detector [335]. Additionally, they applied various detectors, including the MMSE detector, the PDA detector, the LAS detector and the RTS detector, in high-rate non-orthogonal STBC aided LS-MIMO systems [128], [328], [329]. Furthermore, it was shown that non-binary LDPC coded LS-MIMO systems are capable of achieving a near-capacity performance with MMSE detection [336]. It should be noted that in principle a variety of other metaheuristics based MUDs, such as the genetic algorithm (GA) based MUD [529]–[535], the ant colony optimization (ACO) based MUD [537]–[541], the particle swarm optimization (PSO) aided MUD [545]–[547], and the simulated annealing (SA) assisted MUD [548], [549], may also be extended to the LS-MIMO context.

Finally, some soft-input soft-output LS-MIMO detectors having a relatively low complexity were proposed in [338], [339], which rely on the subspace marginalization aided interference suppression (SUMIS) technique and an approximate message passing algorithm, respectively. The first ASIC design of an LS-MIMO detector invoking the truncated Neumann series expansion technique was reported in [340], [341], which achieves a data rate of 3.8 Gb/s for a 3GPP Long Term Evolution-Advanced (LTE-A) based LS-MIMO system having 128 BS antennas and supporting 8 users.

\[ E. \text{Applications of MIMO Detection Techniques in Other Areas} \]

MIMO detection techniques may also be utilized in more advanced scenarios. For example, as a promising technique to utilize the precious radio spectrum more efficiently and
flexibly, cognitive radio (CR) [631]–[633] has stimulated substantial research interests over the past decade. Relying on the software-defined radio (SDR) concept, CR is defined as an intelligent wireless communication system that is capable of learning from the environment and adapting to statistical variations of the environment. Aiming for gaining the benefits of both the CR and MIMO techniques, MIMO cognitive radio has also been studied from various perspectives [634]–[644]. Furthermore, MIMO techniques may also be integrated with the SDR or software-defined network (SDN) for 5G wireless communication systems, where a network function virtualization (NFV) based novel network architecture is envisaged [645].

Apart from their dominant applications in wireless communications, MIMO detection techniques also significantly benefit a range of other research areas. For example, the idea of MIMO signal processing was extended to radar design, and the so-called “MIMO radar”, as illustrated in Fig. 30, has been a hot research topic since the 2000s [646]–[652]. Additionally, MIMO signal processing techniques are also instrumental in mode-division multiplexing (MDM) based multimode fiber (MMF) optical communication systems, as shown in Fig. 31. For more details on MIMO aided high-speed optical communications, please refer to [653]–[659].

X. SUMMARY AND CONCLUSIONS

The concept of LS-MIMO systems may be regarded as a paradigm shift in the wireless communication and signal processing community. In this large dimensional context, the MIMO detection problem becomes even more challenging and important. To facilitate a better understanding of MIMO detection techniques, in this survey, we provided a detailed clarification of the MIMO detection fundamentals, and recited the half-a-century history of MIMO detection. We also provided concise discussions on the distinct detection strategies for different types of LS-MIMO systems and concluded with the recent advances in LS-MIMO detection. Relevant insights and lessons were extracted from the rich heritage of small/medium-scale MIMO detection. We note that when considering the design of LS-MIMO detectors, it is necessary to first identify which type of LS-MIMO system is considered. Specifically, the employment of several popular MIMO detectors, such as the SD based MIMO detectors, may become infeasible in Type-I LS-MIMO systems, while some low-complexity linear MIMO detectors may achieve near-optimum performance in Type-II LS-MIMO systems. Additionally, it was reported that in the LS-MIMO context, local neighborhood search based

metheuristics, Bayesian based message passing methods as well as convex optimization based methods may strike a promising performance versus complexity tradeoff.

ACKNOWLEDGMENT

The authors would like to thank Prof. Hamid Jafarkhani (University of California, Irvine), Prof. Arogyaswami Paulraj (Stanford University) and anonymous reviewers for their insightful comments and suggestions.

REFERENCES


K. A. Singhul, T. Datta, and A. Chockalingam, “Lattice reduction aided detection in large-MIMO systems,” in Proc. IEEE 14th Workshop on


**Shaoshi Yang** (SIEEE’09-MIEEE’13) received his B.Eng. degree in Information Engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China in Jul. 2006, his first Ph.D. degree in Electronics and Electrical Engineering from University of Southampton, U.K. in Dec. 2013, and his second Ph.D. degree in Signal and Information Processing from BUPT in Mar. 2014. He is now working as the IU-ATC Senior Research Fellow in University of Southampton, U.K. From November 2008 to February 2009, he was an Intern Research Fellow with the Communications Technology Lab (CTL), Intel Labs, Beijing, China, where he focused on Channel Quality Indicator Channel (CQICH) design for mobile WiMAX (802.16m) standard. His research interests include MIMO signal processing, green radio, heterogeneous networks, cross-layer interference management, convex optimization and its applications. He has published in excess of 30 research papers on IEEE journals and conferences. Shaoshi has received a number of academic and research awards, including the PMC-Sierra Telecommunications Technology Scholarship at BUPT, the Electronics and Computer Science (ECS) Scholarship of University of Southampton, the Best PhD Thesis Award of BUPT, and the Dean’s Award for Early Career Research Excellence of University of Southampton. He is a member of IEEE/JET, and a junior member of Isaac Newton Institute for Mathematical Sciences, Cambridge University, U.K. He also serves as a TPC member for several major IEEE conferences, including IEEE ICC, PIMRC, ICCVE, HPCC, and as a Guest Associate Editor of IEEE Journal on Selected Areas in Communications. (https://sites.google.com/site/shaoshiyang/)

**Lajos Hanzo** (FREng, FIEEE’04, FIET, FEURASIP, DSc) received his degree in electronics in 1976 and his doctorate in 1983. In 2009 he was awarded the honorary doctorate “Doctor Honoris Causa” by the Technical University of Budapest. During his 37-year career in telecommunications he has held various research and academic posts in Hungary, Germany and the UK. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised 80+ PhD students, co-authored 20 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10 000 pages, published 1400+ research entries at IEEE Xplore, acted both as TPC and General Chair of IEEE conferences, presented keynote lectures and has been awarded a number of distinctions. Currently he is directing a 100-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council’s Advanced Fellow Grant and the Royal Society’s Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also a Governor of the IEEE VTS. During 2008 - 2012 he was the Editor-in-Chief of the IEEE Press and a Chaired Professor also at Tsinghua University, Beijing. Lajos has 20 000+ citations. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk