Spatial Modulation for Generalized MIMO

Citation for published version:

Digital Object Identifier (DOI):
10.1109/JPROC.2013.2287851

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the IEEE

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Spatial Modulation for Generalized MIMO: Challenges, Opportunities, and Implementation

This tutorial paper offers a comprehensive overview of the state of the art in spatial modulation for generalized multiple-input–multiple-output (MIMO) technologies.

By Marco Di Renzo, Member IEEE, Harald Haas, Member IEEE, Ali Ghrayeb, Senior Member IEEE, Shinya Sugiura, Senior Member IEEE, and Lajos Hanzo, Fellow IEEE

ABSTRACT | A key challenge of future mobile communication research is to strike an attractive compromise between wireless network’s area spectral efficiency and energy efficiency. This necessitates a clean-slate approach to wireless system design, embracing the rich body of existing knowledge, especially on multiple-input–multiple-output (MIMO) technologies. This motivates the proposal of an emerging wireless communications concept conceived for single-radio-frequency (RF) large-scale MIMO communications, which is termed as SM. The concept of SM has established itself as a beneficial transmission paradigm, subsuming numerous members of the MIMO system family. The research of SM has reached sufficient maturity to motivate its comparison to state-of-the-art MIMO communications, as well as to inspire its application to other emerging wireless systems such as relay-aided, cooperative, small-cell, optical wireless, and power-efficient communications. Furthermore, it has received sufficient research attention to be implemented in testbeds, and it holds the promise of stimulating further vigorous interdisciplinary research in the years to come. This tutorial paper is intended to offer a comprehensive state-of-the-art survey on SM–MIMO research, to provide a critical appraisal of its potential advantages, and to promote the discussion of its beneficial application areas and their research challenges leading to the analysis of the technological issues associated with the implementation of SM–MIMO. The paper is concluded with the description of the world’s first experimental activities in this vibrant research field.

KEYWORDS | Green and sustainable wireless communications; heterogenous cellular networks; large-scale multiantenna systems; multiantenna wireless systems; network-coded cooperative wireless networks; relay-aided wireless communications; single-radio-frequency (RF) multiantenna systems; spatial modulation; testbed implementation; visible light communications
I. INTRODUCTION AND MOTIVATION: THE ROOTS OF SM

A. Ubiquitous Mobility and Connectivity: The Societal Change

Since the turn of the century, there has been a tremendous growth in the cellular market. The number of
subscribers and the demand for wireless services has escalated. Indeed, the penetration of mobile services exceeded that of the power grid. There are 48 million people in the world who have mobile phones, even though they do not have electricity at home [1].

In this context, mobile communications may be allowed to be an indispensable commodity by most, and mobile data, video as well as television services are also becoming an essential part of everyday life. With the introduction of the Android operating system and the iPhone, the use of ebook readers such as the iPad, and the success of social networking using Facebook, the demand for cellular data traffic has grown significantly in recent years. Thus, communications on the move has proven to be transformational, and mobile operators struggle to satisfy the data traffic demands in wireless cellular networks, while keeping their costs at minimum to maintain profitability.

B. Mobile Data Traffic: The Forecast

A further explosion of mobile data traffic is predicted. According to Cisco’s estimates predicted in February 2011 [1], the 2010 mobile data traffic growth rate was higher than anticipated. The global mobile data traffic grew 2.6-fold in 2010, nearly tripling for the third year in a row. Furthermore, according to Cisco’s estimates published in February 2013 [2], the global mobile data traffic grew 70% in 2012, and it reached 885 petabytes (PB) per month at the end of 2012, up from 520 PB per month at the end of 2011. These growth rates of mobile data traffic resemble those of the fixed network observed during 1997–2001, when the average yearly growth was 150%. The overall mobile data traffic is expected to grow to 11.2 exabytes (EB) per month by 2017, a 13-fold increase over 2012, which corresponds to a CAGR of 66% from 2012 to 2017. More particularly, the Asia Pacific and North America regions will account for almost two-thirds of the global mobile traffic by 2017. Middle East and Africa will experience the highest CAGR of 77%, increasing 17.3-fold over the forecast period. The Asia Pacific region will have the second highest CAGR of 76%, increasing 16.9-fold over the forecast period. The emerging regions of Latin America as well as Central and Eastern Europe will have CAGRs of 67% and 66%, respectively. When combined with the Middle East and Africa, the aforementioned emerging market will represent an increasing share of the total mobile data traffic, which is expected to be up from 19% at the end of 2012 to 22% by 2017 [2].

C. Mobile Data Traffic: The In-Home Scenario

Furthermore, a survey conducted by Cisco IBSG indicates that much of the mobile data activity takes place in the home. In particular, it has been estimated that the percentage of time spent using the mobile Internet at home is approximately 40%. The amount of mobile data traffic on the move is approximately 35%, while the remaining 25% of mobile Internet use occurs at work. The relatively high percentage of home-based mobile data use suggests that next-generation cellular networks require specific data access points installed by home users to satisfy the huge demand for data traffic, and, at the same time, to get improved indoor voice and data coverage. By using these home access points, the telecommunication operators may be able to offload, in a cost- and energy-effective manner, the data traffic onto a fixed network, either by offering their subscribers dual-mode mobile phones or through the employment of femtocells, which are considered the key enabling technology to handle the growing demands for mobile data traffic in the home [3]. In particular, to meet the demand of massive mobile data growth, IDATE Research & Consulting and Infonetics Research has forecast the employment of 39.4 million femtocell units and a $2.98 billion market by 2015 [4].

D. Next-Generation Cellular Networks: The Compelling Need to be “Green”

The unprecedented surge of mobile data traffic in the cellular industry has motivated telecommunications operators and researchers to develop new transmission technologies, protocols, and network infrastructure solutions for maximizing both the achievable throughput and the spectral efficiency. On the other hand, little or no attention has been devoted to energy consumption and complexity issues. As a result, the ICT sector contributes substantially to the global carbon emissions.

In particular, at the time of writing, the ICT sector represents around 2% of the global carbon emissions already, of which mobile networks contribute about 0.2%. This is comparable to the worldwide carbon emissions of airplanes, and about a quarter of the worldwide carbon emissions of cars. Furthermore, this amount is expected to increase every year at a rapid pace due to the massive increase of the mobile data traffic. Currently, there are more than 5 million BSs serving mobile users, each consuming an average of 25 MWh per year [5], [6]. In addition to the environmental aspects, the energy costs represent a significant portion of the network operators OPEX. While each BS connected to the electrical grid may cost approximately $3000 per year to operate, off-grid BSs operating in remote areas generally run on diesel power generators and may cost ten times more [5]. Furthermore, with the advent of data-intensive cellular standards, such as the LTE-A system, the energy consumption of each BS can increase up to 1400 W, and the energy cost of each BS may reach $3200 per annum with a carbon footprint of 11 tons of carbon emissions [7]. The radio network itself adds up to 80% of an operator’s entire energy consumption.

In this context, the development of revolutionary clean-slate wireless communications technologies that are capable of meeting the forecast mobile data traffic growth while reducing the carbon footprint of next-generation cellular networks is a compelling necessity.
E. Addressing the Energy-Efficiency Challenge: Green Small-Cell Assisted Networks

The rising energy cost and carbon footprint of operational cellular networks have motivated both network operators and regulatory bodies, such as the 3GPP and the ITU, to develop innovative solutions for improving the energy efficiency of cellular systems. This emerging trend has attracted the interest of researchers worldwide to develop “green small-cell assisted networks” [8].

Numerous collaborative projects have been launched worldwide for addressing the energy efficiency of mobile communications systems. Notable examples are:

- the Energy Aware Radio and NetWork TecHnologies (EARTH) project [9];
- the Towards Real Energy-efficient Network Design (TREND) project [10];
- the Cognitive Radio and Cooperative strategies for Power saving in multi-standard wireless devices (C2POWER) project [11];
- the GREENET—An early stage training network in enabling technologies for green radio project [12];
- the Green Terminals for next generation wireless systems (GREEN-T) project [13];
- the GreenTouch consortium [14], whose mission is to deliver the architecture, the specifications, and the roadmap to increase, by 2015, the network’s energy efficiency by a factor of 1000 compared to the 2010 levels.

Furthermore, in recent press releases (e.g., IP/09/393 [15]), ICT players have been warmly invited to develop innovative technologies in support of a greener world and to make people more aware of how they use energy. In this context, “Green Radio” constitutes a wide-ranging research discipline that intends to cover all layers of the protocol stack and various system architectures, as well as to identify the fundamental tradeoffs between energy efficiency and system-wide performance.

F. The Emerging Paradigm Shift: From SE to EE

The conventional response to the surge of mobile data traffic is the proposal of advanced transmission technologies and protocols designed for maximizing the SE. In fact, since the SE is directly linked to the notion of Shannon capacity [16], until recently it has been considered to be the main performance indicator fueling the design and optimization of wireless communications systems in general and cellular networks in particular. As a result, the vast majority of transmission technologies and protocols used in the operational cellular and mobile networks have been designed by taking into account diverse factors, such as throughput, QoS, availability, scalability, etc., without paying specific attention to the energy consumption. With this design methodology, the operational cellular systems can only achieve energy savings at the cost of a performance and/or throughput degradation. Explicitly, it is crucial to develop power-efficient, low-complexity solutions that still satisfy the target QoS and throughput requirements.

To this end, transmission technologies and protocols should be designed and optimized for next-generation cellular networks by using more appropriate performance indicators, which explicitly take the energy consumption and the system’s complexity into account. A metric that is often used to this end is the EE, which provides an indication of the throughput per unit energy [16], [17]. When using this metric for system design, the first important conclusion is that current solutions that are spectral efficient turn out to be suboptimal in terms of energy efficiency [16], [18]. More specifically, the EE metric decreases monotonically with the throughput, which leads to the conclusion that energy-efficient solutions are expected to operate relatively far from the Shannon capacity. This fundamental tradeoff emerges even for the basic point-to-point AWGN channel, where the SE versus EE tradeoff may be formulated as

\[ \eta_{EE} = \eta_{SE} N_0 (2^{\eta_{SE}} - 1)^{-1} \]

with \( N_0 \) being the receiver noise power spectral density, while \( \eta_{SE} \) and \( \eta_{EE} \) denote the SE and the EE, respectively. This simple formula highlights that the EE is monotonically decreasing when increasing the SE. However, improving the EE at the cost of the QoS (SE/throughput) for the end user may be unacceptable in commercial networks.

In summary, the development of beneficial wireless communications techniques striking an attractive SE versus EE tradeoff for next-generation cellular networks is a compelling necessity [18].

II. SINGLE-RF LARGE-SCALE SM–MIMO: AN IDEA WHOSE TIME HAS COME

A. The Cellular Network of the Future

The traditional cellular network fails to keep pace with the mobile data explosion forecasts. We need innovative technologies and cellular topologies that can meet these demands in an energy-efficient and sustainable manner. To address this challenge, and, thereby, to maintain profitability, it is crucial to develop energy-efficient wireless architectures, transmissions schemes, protocols, and cooperative relaying and heterogeneous network solutions based on smaller cells [3], [8].

More specifically, competing forces, i.e., spectral efficiency and throughput versus energy efficiency and low complexity, are rapidly changing the topology of operational cellular networks, which are undergoing a major change: the migration from voice-centric, circuit-switched, and centrally optimized networks toward data-centric, packet-switched, and high-throughput networks [3]. The cellular network of the future will be:

- heterogeneous and characterized by a small-cell infrastructure relying on inexpensive and low-power BSs (femtocells) in order to achieve high data rates [3], [8];
• green, by evolving from a throughput-optimized scenario toward throughput- and energy-optimized networks [5], [16], [18];
• interference aware, by exploiting (instead of tolerating) interference and, thus, realizing the expected benefits of small-cell-based heterogeneous networking [19];
• characterized by a high level of cooperation among BSs and user terminals, by achieving improved coverage and reduced energy consumption through relay-aided transmission, as well as an improved reliability and reduced packet transmissions/retransmissions through distributed diversity and NC, respectively [20]–[22];
• relying upon new air-interface techniques and physical-layer standards for increasing the EE, reducing the implementation and signal processing complexity, while meeting the required SE. This important issue is extensively addressed in this paper.

The advantages of the emerging heterogeneous cellular network architecture are numerous. Heterogeneity expands the coverage, improves the network capacity, reduces the energy consumption, and enhances the link reliability through a more dense deployment of low-cost and low-power access points. The reason behind all these potential advantages is simple: the densification of access points inherently reduces the distance between the network elements. Since, based on electromagnetic laws, the received power falls off exponentially with the transmission distance, this implies that reducing the distance has a beneficial impact on both the achievable capacity and on the transmission power. In addition, unlike macro BSs that are usually heavily loaded most of time, the femto BSs are expected to be lightly loaded, which leads to further reduction of the power consumption since they will not be transmitting all the time, hence facilitating sleep-mode-based operation.

B. A Physical-Layer Perspective

It is widely recognized that the two most promising physical-layer standards of 4G cellular networks are the mobile WiMAX release 2 based on the 802.16m-2011 protocol [23], [24], and, especially, LTE-A [25], [26]. The power efficiency of the ICT industry in the next decade will highly depend on the EE of these physical-layer standards. However, at the current stage, both standards may be deemed to be conceived, designed, and optimized based on the SE, with limited [27] consideration of the EE issues. In fact, especially at the physical layer, the primary focus has been on achieving high data rates, without giving much cognizance to the EE and implementation complexity. However, this approach is no longer applicable to future cellular networks.

Both the WiMAX and LTE-A physical-layer standards heavily rely on MIMO technology for realizing their achievable throughput [28], [29]. MIMO communications constitute promising techniques for the design of future wireless communications systems, including the 5G cellular networks. In simple terms, the capacity of MIMO systems is proportional to \( \min\{N_t, N_r\} \) under some benign propagation conditions [29, eq. (1.17)], where \( N_t \) and \( N_r \) represent the number of TAs and RAs. In particular, this theoretical limit can be achieved if channel side information is available at both the transmitter and the receiver, the transmit-to-receive links are sufficiently independent, as well as the SNR is sufficiently high. Under these favorable conditions, the throughput may be increased linearly with the number of antennas. As a consequence, MIMO techniques are capable of providing high data rates without increasing the spectrum utilization and the transmit power.

However, in practice, MIMO systems need a multiplicity of associated circuits, such as power amplifiers, RF chains, mixers, synthesizers, filters, etc., which substantially increase the circuit power dissipation of the BSs [5], [30]–[32]. More explicitly, recent studies have clearly shown that the EE gain of MIMO communications increases with the number of antennas, provided that only the transmit power of the BSs is taken into account and their circuit power dissipation is neglected. On the other hand, the EE gain of MIMO communications remains modest and decreases with the number of active TAs, if realistic power consumption models are considered for the BSs [33]. These results highlight that the design of EE–MIMO communications conceived for multiantenna multicell networks is a fairly open research problem, where many system parameters have to be considered, such as the bandwidth, the transmit power, the number of active TAs/RAs, the number of active users, etc., which all contribute to the fundamental transmit power versus circuit power dissipation and multiplexing gain versus interuser interference tradeoffs [34].

As a result, while the SE advantages of MIMO communications are widely recognized, the EE potential of MIMO communications for cellular networks is not well understood. For example, recent results have shown that, under a total consumed power optimization constraint for the power amplifiers, the TAs with the weakest channel gains should be turned off [35]. This allows the BSs to turn off the related RF chains with filters and mixers, which saves additional power. Furthermore, the signal processing complexity of optimal signal detection algorithms to be used at the mobile terminals makes the practical implementation of MIMO systems a challenge [28], [36]. Hence, new air-interface transmission techniques have to be developed that are capable of striking an attractive tradeoff between the SE and the EE, rather than aiming for SE optimization only.

As far as the physical-layer design is concerned, radical new transmission techniques have to be conceived considering both EE and SE optimization.
C. From MIMO to SM–MIMO

Conventional MIMO communications take advantage of all the antennas available at the transmitter by simultaneously transmitting multiple data streams from all of them. Thus, all TAs are active at any time instance. By appropriately choosing the transmission/precoding matrices, both multiplexing and transmit-diversity gains can be obtained using MIMOs [28]. The reason behind this choice is that simultaneously activating all TAs results in SE optimization [37]. Unfortunately, this choice does not lead to EE optimization [34], [37]–[39]. For example, in [33], it is shown that, under realistic BS power consumption models, MIMO systems equipped with more than two active TAs unlikely provide any total EE gains at the current state of the art.

Compared to baseline single-antenna transmissions, MIMO communications obtain higher data rates and improved error performance at the cost of:

- increasing the signal processing complexity at the receiver, which is caused by the need for counteracting the interference imposed by simultaneously transmitting many data streams;
- more stringent synchronization requirements among the TAs to exploit the benefits of space-time-coded and multiuser MIMO transmissions;
- multiple RF chains at the transmitter to be able to simultaneously transmit many data streams, which do not scale with Moore’s law and make the transmitter bulky [36];
- independent power amplifiers for each RF chain, which dissipate the majority of the power consumed at the transmitter, since they are power inefficient due to the stringent linearity requirements of state-of-the-art phase/amplitude modulations [5].

These considerations imply that a major challenge of next-generation MIMO-aided cellular networks is the design of multiantenna transmission schemes with a limited number of active RF chains aiming for reducing the complexity, to relax the interantenna synchronization requirements, and ICI, as well as the signal processing complexity at the receiver, while aiming for improving the EE.

In this context, single-RF MIMO design is currently emerging as a promising research field [36]. The fundamental idea behind single-RF MIMO is to realize the gains of MIMO communications, i.e., spatial multiplexing and transmit diversity, with the aid of many antenna elements, of which only a few, possibly a single, activated antenna elements (single-RF front–end) at the transmitter at any modulation instant [40]. The rationale behind the multi-RF to single-RF paradigm shift in MIMO design originates from the consideration that large numbers of TAs (radiating elements) may be accommodated at the BSs (large-scale MIMO design) [41], [42], especially in the emerging millimeter-wave band [43]–[46], bearing in mind that the complexity and power consumption/dissipation of MIMO communications are mainly determined by the number of simultaneously active TAs, i.e., by the number of active RF chains [40], [47].

Fueled by these considerations, SM has recently established itself as a promising transmission concept, which belongs to the single-RF [36] large-scale [41] MIMO wireless systems family, while exploiting the multiple antennas in a novel fashion compared to state-of-the-art high-complexity and power-hungry classic MIMOs [48]. In simple terms, SM can be regarded as a MIMO concept that possesses a larger set of radiating elements than the number of transmit electronics. SM–MIMO takes advantage of the whole antenna array at the transmitter, while using a limited number of RF chains. The main distinguishing feature of SM–MIMOs is that they map additional information bits onto an “SM constellation diagram,” where each constellation element is constituted by either one or a subset of antenna elements. These unique characteristics facilitate high-rate MIMO implementations to have reduced signal processing and circuitry complexity, as well as an improved EE [49]. Recent analytical and simulation studies have shown that SM–MIMOs have the inherent potential of outperforming many state-of-the-art MIMO schemes, provided that a sufficiently large number of antenna elements is available at the transmitter, while just few of them are simultaneously active [50].

In a nutshell, the rationale behind SM–MIMO communications design for spectral- and energy-efficient cellular networks is centered upon two main pillars:

1) given the performance constraints, minimize the number of active antenna elements in order to increase the EE by reducing the circuit power consumption at the transmitter (single-RF MIMO principle);

2) given the implementation and size constraints, maximize the number of passive antenna elements in order to increase both the SE and the EE by reducing the transmit power consumption (large-scale MIMO principle). This is realized by capitalizing on the multiplexing gain introduced by mapping additional bits onto the “SM constellation diagram.”

D. Paper Organization

This paper is intended to offer a comprehensive state-of-the-art survey on SM–MIMO research, a critical appraisal of its beneficial application areas and their research challenges, an analysis of the technological issues associated with the implementation of SM–MIMO, and, finally, a description of the world’s first experimental activities in this research field.

The remainder of this paper is divided in five interlinked parts.

1) We commence by providing a detailed description of the operating principle of SM–MIMO, the analysis of its advantages and disadvantages
TRANSCIVER DESIGN

A. SM–MIMO: How It Works

In this section, we commence by introducing the SM–MIMO concept, illustrating it with the aid of some simple examples. Again, we denote by $M$ the number of TAs and RAs, respectively. The cardinality of the signal-constellation diagram is denoted by $M$. Either PSK or QAM are considered. In general, $N_t$, $N_r$, and $M$ can be chosen independently of each other. At the receiver, optimum ML demodulation is considered. Thus, $N_t$ can be chosen independently of $N_t$ [28]. For ease of exposition, we assume $N_t = 2^n$ and $M = 2^m$ with $n$ and $m$ being two positive integers. In Section IV, we describe general SM–MIMO encodings as well as some suboptimal (non-ML) demodulation schemes.

In Fig. 1, the SM–MIMO concept is illustrated for $N_t = M = 2$, and it is compared to the conventional SMX scheme [28, Sec. II] and the OSTBC scheme designed for transmit diversity [28, Sec. III]. In the latter case, the Alamouti scheme is considered as an example [51].

1) In SMX–MIMO, two PSK/QAM symbols ($S_1$ and $S_2$) are simultaneously transmitted from a pair of TAs in a single channel use. For arbitrary $N_t$ and $M$, the rate of SMX is $R_{SMX} = N_t \log_2(M)$ bpcu [28, Sec. II].

2) In OSTBC–MIMO, two PSK/QAM symbols ($S_1$ and $S_2$) are first encoded and then simultaneously transmitted from a pair of TAs in two channel uses. For arbitrary $N_t$ and $M$, the rate of OSTBC is $R_{OSTBC} = R_c \log_2(S) \text{ bpcu}$, where $R_c = N_t/M/N_{cu} \leq 1$ is the rate of the space-time block code and $N_{M}$ is the number of information symbols transmitted in $N_{cu}$ channel uses [28, Sec. III], [52]. If, as shown in Fig. 1, the Alamouti code is chosen, then we have $R_c = 1$.

3) In SM–MIMO, only one ($S_1$) out of the two symbols is explicitly transmitted, while the other symbol ($S_2$) is implicitly transmitted by determining the index of the active TA in each channel use. In other words, in SM–MIMO, the information symbols are modulated onto two information carrying units: a) one PSK/QAM symbol; and b) a single active TA via an information-driven antenna-switching mechanism. For arbitrary $N_t$ and $M$, the rate of SM is $R_{SM} = \log_2(M) + \log_2(N_t/\text{bpcu})$ [48], [53].

In Figs. 2 and 3, the encoding mechanism of SM–MIMO is illustrated for $N_t = M = 4$ by considering two generic channel uses, where the concept of “SM or spatial-constellation diagram” is also introduced. The rate of this MIMO setup is $R_{SM} = \log_2(M) + \log_2(N_t) = 4$ bpcu, hence the encoder processes the information bits in blocks of four bits each. In the first channel use shown in Fig. 2, the block of bits to be encoded is “1100.” The first $\log_2(N_t) = 2$ bits, “11,” determine the single active TA (TX$_1$), while the second $\log_2(M) = 2$ bits, “00,” determine the transmitted PSK/QAM symbol. Likewise, in the
second channel use shown in Fig. 3, the block of bits to be encoded is “0001.” The first \( \log_2(N_t) = 2 \) bits, “00,” determine the single active TA (TX0), while the second \( \log_2(M) = 2 \) bits, “01,” determine the transmitted PSK/QAM symbol.

The illustrations shown in Figs. 2 and 3 highlight a pair of unique characteristics of SM–MIMO.

1) The activated TA may change every channel use according to the input information bits. Thus, TA switching is an effective way of mapping the information bits to TA indices and of increasing the transmission rate. It is worth mentioning here that the idea of increasing the rate of wireless communications using TA switching has been alluded in pioneering MIMO papers under the concept of “spatial cycling using one transmitter at a time” [54, p. 317].

2) The information bits are modulated onto a 3-D constellation diagram, which generalizes the known 2-D (complex) signal-constellation diagram of PSK/QAM modulation schemes. The third dimension is provided by the antenna array, where some of the bits are mapped to the TAs. In SM–MIMO research, this third dimension is termed the “spatial-constellation diagram” [48].

In simple mathematical terms, the signal model of SM–MIMO, assuming a frequency-flat channel model, is as follows:

\[
y = Hx + n
\]  

where \( y \in \mathbb{C}^{N_t \times 1} \) is the complex received vector; \( H \in \mathbb{C}^{N_r \times N_t} \) is the complex channel matrix; \( n \in \mathbb{C}^{N_r \times 1} \) is the complex AWGN at the receiver; and \( x = es \in \mathbb{C}^{N_t \times 1} \) is the complex modulated vector with \( s \in \mathcal{M} \subseteq \mathbb{C}^{1 \times 1} \) being the complex (scalar) PSK/QAM modulated symbol belonging to the signal-constellation diagram and \( e \in \mathcal{A} \) being the \( N_t \times 1 \) vector belonging to the spatial-constellation diagram \( \mathcal{A} \) as follows:

\[
e_t = \begin{cases} 1, & \text{if the } t\text{th TA is active} \\ 0, & \text{if the } t\text{th TA is not active} \end{cases}
\]  

where \( e_t \) is the \( t\)th entry of \( e \) for \( t = 1, 2, \ldots, N_t \). In other words, the points (\( N_t\)-dimensional vectors) of the spatial-constellation diagram are the \( N_t \) unit vectors of the natural basis of the \( N_t\)-dimensional Euclidean space.

If \( N_t = 1 \), SM–MIMO reduces to conventional single-antenna communications, where the information bits are encoded only onto the signal-constellation diagram. In this case, the rate is \( R_0 = \log_2(M) \). On the other hand, if \( M = 1 \) the information is encoded only onto the spatial-constellation diagram by providing a rate equal to \( R_{SSK} = \log_2(N_t) \). In the literature, this transmission scheme is known as SSK modulation [55], which is detailed in Section IV. In particular, SSK modulation is a MIMO scheme, where data transmission takes place only through the information-driven TA switching mechanism. It is apparent that SM–MIMO can be viewed as the combination of single-antenna PSK/QAM and SSK–MIMO modulations.

B. SM–MIMO: Why It Works

From Section III-A, we have learned that the fundamental operational principle of SM–MIMO relies on transmitting part of the information bits via an implicit information-driven antenna-switching mechanism. In this section, we provide some insights on how the receiver becomes capable of retrieving the information bits encoded into the TA indices. This, in fact, showcases the underlying “essence” that equips SM–MIMO with its innate properties. Let us consider Figs. 4–6, which illustrate the bit-to-symbol mapping and the transmission process through the communication channel as well as the decoding process of SM–MIMO transmission.
respectively. Fig. 4 is similar to Figs. 2 and 3, but \( N_t = 4 \) and \( M = 2 \) are considered for ease of illustration. Let “101” be the triplet of bits at the input of the SM modulator. Then, the TA element \( TX_2 \) is activated and the “−1” PSK bit is transmitted from it.

The signal emitted by the active TA then traverses through a generic wireless channel, whose channel impulse responses are illustrated in Fig. 5. Owing to the different spatial locations occupied by the TAs in the antenna array, the signal transmitted by the active TA experiences different propagation conditions due to the different interacting environmental objects along any transmitter-to-receiver wireless links. As such, the same “−1” PSK symbol emitted by the TA element \( TX_2 \) travels through a communication channel, which introduces a specific “channel signature or fingerprint,” i.e., the channel impulse response, that makes it unique compared to the same symbol emitted by any other TAs. This constitutes the fundamental essence of SM–MIMO: the more different the channel signatures/fingerprints are from each other, the simpler it becomes to distinguish the signals at the receiver. In the following, we refer to this condition as “favorable propagation conditions.” In simple terms, the communication channel may be deemed to play, especially for SSK–MIMO, the role of a “modulation unit,” where the channel impulse responses are the actual messages being transmitted.

At the receiver, the demodulation unit exploits the unique fingerprint introduced by the wireless channel for retrieving the information bits. This is illustrated in Fig. 6, where a coherent demodulation scheme based on the minimum Euclidean distance is considered [56]. The receiver is assumed to be aware of the \( N_t \) channel impulse responses, however the actual channel impulse response that is received in each channel use depends on the index of the active TA. The demodulator performs an exhaustive search among all the possible combinations of channel impulse responses and modulation symbols, and makes a decision in favor of the hypothesis associated with the lowest Euclidean distance. In a nutshell, due to the information-driven antenna-switching mechanism of SM–MIMO transmission, the \( N_t \) channel impulse responses become part of the search space of the hypothesis-testing problem solved by the receiver. Based on the estimated channel impulse response, the demodulator is capable of retrieving the information bits associated with it. In summary, the essence of SM–MIMO transmission is all about exploiting the TA-specific property of the wireless channel, i.e., the uniqueness of each transmit-to-receive wireless link, for data communication.

### C. Potential Advantages and Disadvantages of SM–MIMO

In the light of the encoding principle described in Section III-A, SM–MIMO provides the following potential advantages compared to state-of-the-art MIMO communications.

- **Higher throughput.** Thanks to the 3-D constellation diagram and to the introduction of the spatial constellation, the SE of SM–MIMO is higher than that of single-antenna and OSTBC–MIMO transmission. This improved SE (MIMO gain) translates into the reduction of the RF output power [28].
• **Simpler receiver design.** Since only a single TA is active in every channel use, SM–MIMO is not affected by ICI, hence it provides ML-optimum performance at a single-stream decoding complexity [56].

• **Simpler transmitter design.** Due to the single TA transmission, SM–MIMO can be implemented by using a single active RF chain and many inactivated TAs, which is inexpensive and easy to deploy [36]. Thus, the employment of multiple expensive and bulky power amplifiers, RF filters/mixers, analog-to-digital converters, and RF coaxial cables can be avoided. For example, the typical RF coaxial cables used in current tower-mounted BSs are more than four centimeters in diameter [42].

• **Lower transmit power supply.** Since the multiplexing gain is achieved by a single-RF source, SM–MIMO reduces the total consumed power required for the same RF output power. In particular, the power dissipation is independent of the number of TAs. This results in EE gains, especially for low/medium and medium/high SEs [33].

• **Better efficiency of the power amplifiers.** The power efficiency of current power amplifiers decreases when increasing the linearity requirements of the modulation scheme (e.g., QAM) [5]. Recent results have shown that SM–MIMO associated with constant-envelope modulation (e.g., PSK modulation) is capable of providing similar or even better performance than amplitude modulation schemes (e.g., QAM) [57]. The reduced linearity requirements of constant-envelope modulation increase the efficiency of power amplifiers, which, in turn, reduce the total power consumption of the transmitters [5]. This power efficiency can be further improved if SSK modulation is used.

These important advantages of SM–MIMO introduce, however, some fundamental tradeoffs.

• **SE suboptimality.** Since some TA elements remain inactive in every channel use, SM–MIMO offers a lower throughput than SMX–MIMO. In particular, $R_{SMX}$ increases linearly with $N_t$, while $R_{SM}$ increases logarithmically with $N_t$. This implies that SM–MIMO requires a larger number of TA elements for achieving the same SE as capacity-achieving SMX–MIMO communications. This potential disadvantage may, however, be offset by the emerging large-scale MIMO and millimeter-wave cellular communications paradigms [41]–[46], which foresee the future employment of compact transmitters and receivers having hundreds of TA elements. For example, in the 28-GHz band, the size of a $12 \times 4 = 48$ horn antenna array occupies only $11.66 \text{ cm} \times 6.81 \text{ cm}$ in area [45]. In this context, however, important practical issues, such as the training overhead for channel estimation and feedback for large-scale MIMO implementations and the need for directional beamforming gains for application to millimeter-wave frequencies, need to be carefully investigated, as described in this section.

• **Fast antenna switching.** Due to its specific encoding mechanism, in SM–MIMO, the active TA changes in every channel use. As a consequence, a single-RF implementation needs a sufficiently fast RF switch operating at the symbol rate that introduces low insertion/switching losses. Thus, high-speed RF switches constitute a critical part of the transmitter design. Fortunately, several examples of RF switches capable of switching at nanosecond or even at subnanosecond speeds with low insertion loss and with good isolation properties are available in the literature for a wide range of frequency bands [58]–[61].

• **Time-limited pulse shaping.** The operational wireless communication standards require the transmitted signals to obey a well-defined spectral mask. More specifically, the transmitted signals are usually designed to have a flat spectrum in the transmission band for improving the energy efficiency, as well as a fast rolloff in order to reduce the out-of-band interference and to enhance the associated coexistence capabilities. To this end, appropriate shaping filters have to be used before transmission. Commonly adopted shaping filters satisfy the Nyquist criterion, thus they are bandwidth limited and hence have an infinite-duration impulse response. On the other hand, the SM–MIMO encoding based on a symbol-time switching mechanism is better suited for time-limited pulse shapes [55], which results in a bandwidth expansion. Thus, in SM–MIMO systems, pulse shaping should be realized with the objective of striking a good tradeoff between a limited time duration and a practical bandwidth occupancy in order to ease the implementation of the switching mechanism, while meeting the practical coexistence requirements. Fortunately, various methods are available in the literature for generating practical time- and bandwidth-limited pulse-shaping filters [62].

• **Favorable propagation conditions.** The efficiency of SM–MIMO communications depends on the radio environment, which is termed in the large-scale MIMO literature as *favorable propagation* [63]. Briefly, this implies that the channel impulse responses of the transmit-to-receive links are sufficiently different from each other [48]. In fact, the channel impulse responses represent the unique points/signatures of the spatial-constellation diagram, which implies that the more different they
are, the easier for the demodulator to distinguish them is. This leads to an improved error performance and EE. As a consequence, the lack of scattering in the propagation environment may result in a poor error probability and EE. However, adequate transmit preprocessing, such as orthogonal pulse shaping, is capable of alleviating these problems [64]. Furthermore, the routinely encountered practical design issues of conventional MIMO communications, such as RF power imbalance, may no longer be a serious problem in SM–MIMO, since they in fact assist in making the channel impulse responses more different from each other [66].

- **Training overhead.** The achievable throughput of SM–MIMO systems with a single RF chain increases logarithmically with the number of TAs. Furthermore, even though conventional SM–MIMO schemes are open loop and do not need the knowledge of the CSIT, the receiver still needs to estimate the channel impulse responses of all the TAs for ML-optimum demodulation. As a consequence, large-scale SM–MIMO implementations may incur in a nonnegligible training overhead for channel estimation. Moreover, the constraint (for complexity reduction and EE improvement) on a limited number of RF chains may increase further the training overhead, since the channel gains of all the TAs cannot be estimated simultaneously. The assessment of the potential gains of SM–MIMO schemes compared to conventional MIMO schemes need to accurately take into account this training overhead [65]. If closed-loop SM–MIMO schemes are considered (see Section IV-B), the feedback overhead for CSIT reporting needs to be taken into account as well. These important aspects for the design of SM–MIMO systems are discussed in Section IV-F and I, and more in detail in Section VI-B.

- **Directional beamforming.** To achieve the same throughput as conventional SMX–MIMO systems with a reduced number of RF chains, SM–MIMO systems need large-scale implementations with tens/hundreds of TAs. The emerging millimeter-wave frequency band for cellular communications offers the opportunity of compactly packing many TAs at both the transmitter and the receiver, thanks to the decrease in wavelength. Thus, SM–MIMO communications may be deemed to be suitable for application to millimeter-wave frequencies. In this frequency band, however, directional beamforming is a compelling necessity in order to overcome the higher pathloss and the higher noise level compared to the frequency bands currently used by operational cellular systems. As a consequence, for application to the millimeter-wave frequency band, SM–MIMO systems need to simultaneously provide a multiplexing gain and a directional beamforming gain, while retaining their low-complexity implementation. This important issue is discussed in detail in Section VI-D.

### D. Generalized MIMO Transceiver Based on SM

The ratification of widely accepted new wireless communication standards is typically based on years of research by standardization bodies. In that regard, SM–MIMO is a new concept, which would require major changes in the state-of-the-art air-interface standards, such as the LTE-A standard, where, at most, eight and four antenna ports are available at the BS and at the mobile terminal, respectively [26]. However, we believe that SM–MIMO should not necessarily be considered as a competing MIMO technique in the LTE-A standard, but, rather, as an enhancement toward spectral- and energy-efficient 5G cellular networks.

To put forth this view, in this section, we introduce a generalized MIMO transceiver based on the SM principle, and we show that it may be harmonized with the fundamental transmission modes of the LTE-A standard. More specifically, by “switching off” the spatial-constellation diagram, the proposed SM–MIMO transceiver reduces to the LTE-A space-time-coded transmission. On the other hand, by “switching on” the spatial-constellation diagram, we can improve the throughput, the error performance, and/or the power efficiency without imposing and excessive complexity to the transmitter and to the receiver. The interested reader can find further details about this generalized SM–MIMO transceiver in [50].

1) **Signal Model:** Consider a generic MIMO system associated with $N_t$ and $N_r$ antenna elements at the transmitter and at the receiver, respectively. Transmission occurs over a frame duration of $N_s$ time slots. The time-slot duration is denoted by $T_s$. The channel is assumed to be flat fading and the complex path gain from the $t$th TA to the $r$th RA is denoted by $H_{t,r}$. These path gains are collected in an $(N_r \times N_t)$-element matrix denoted by $H$. The wireless channel is assumed to be quasi-static during the transmission frame, so that the path gains remain constant over $N_s$ time slots, while they change independently from one transmission frame to another.

The following notation is used. Matrices and column vectors are shown in boldface uppercase and lowercase letters, respectively. The $i$th entry of vector $x$ is denoted by $x_i$. The $(i,j)$th entry of matrix $X$ is denoted by $X_{ij}$. Transpose, Hermitian, complex conjugate, and square absolute values are denoted by $(\cdot)^T$, $(\cdot)^H$, $(\cdot)^*$, and $|\cdot|^2$, respectively. The Frobenius norm is denoted by $\| \cdot \|_F$. The cardinality of a set is denoted by $\text{card} \{ \cdot \}$. The operator $x \mod y = \text{mod}(x - 1, y) + 1$ is introduced, where $x$ and $y$ are positive integer numbers and $\text{mod}(\cdot, \cdot)$ is the remainder operator.
The signal received at the rth RA and at the s-th time slot is given in \((s-1)T_s \leq \xi < sT_s\)

\[
z_{r,s}(\xi) = \begin{cases} 
\sqrt{E_s} \sum_{\ell=1}^{N_s} \|a^{(\ell)}\|_F^2 \left[ X_{s,t}^{(\ell)}(\mu)H_{r,t}w_t(\xi) \right] + n_{r,s}(\xi), & \text{if } \|a^{(\ell)}\|_F^2 \neq 0 \\
n_{r,s}(\xi), & \text{if } \|a^{(\ell)}\|_F^2 = 0 
\end{cases}
\]

(3)

where we have defined

\[
X_{s,t}^{(\ell)}(\mu) = a_{t,\ell}^{(\ell)} M_{s,t} \mu
\]

(4)

and \(a^{(\ell)}\) is an \((N_t \times 1)\)-element vector whose \(t\)-entry is \(a_{t,\ell}^{(\ell)} \in \{0, 1\}\). It identifies the TAs that are active over the frame duration of \(N_t\) time slots. The number of active antennas in \(a^{(\ell)}\) is equal to the number of its nonzero entries, i.e., \(N_{a,\ell} = \|a^{(\ell)}\|_1^2\). The ensemble of vectors \(a^{(\ell)}\) is the spatial-constellation diagram, which is denoted by \(A = \{a^{(\ell)}|\alpha = 1, 2, \ldots, A\}\), where \(A = \text{card}\{A\}\). \(M(\cdot)\) is the \(N_t \times N_t\) transmission matrix of a mother space-time code, which uses the \(N_t\) active TAs over the frame duration of \(N_t\) time slots. The entries of \(M(\mu)\) contain specific linear combinations of the entries of \(\mu = [\mu_1, \mu_2, \ldots, \mu_{NM}]^T\), where each element of \(\mu\) is a multilevel amplitude- and phase-modulated complex symbol, e.g., PSK/QAM, and \(NM\) is the number of information symbols transmitted from the \(N_{a,\ell}\) active TAs over the frame duration of \(N_t\) time slots [28], [51], [52]. The set \(M(\cdot)\) of \(M = \text{card}\{M\}\) complex symbols is the signal-constellation diagram, where \(\mu_m \in M(\cdot)\) is the shaping filter at the \(t\)-th TA before transmission. The shaping filters are assumed to be real-valued functions with unit energy and a finite duration, which is equal to the time slot \(T_s\). \(E_s\) denotes the average transmitted energy per time slot (channel use), and the normalization by \(\|a^{(\ell)}\|_F^2\) ensures the total energy constraint is obeyed by the active TAs. Since no channel state information is considered at the transmitter, \(E_s\) is equally split among the active TAs. \(n_{r,s}(\cdot)\) is the AWGN at the input of the rth RA and at the s-th time slot. The noise samples are circularly symmetric complex Gaussian random variables with zero mean and variance \(N_0\) per real dimension, after filtering with \(w_t(\cdot)\). The noise samples across the RAs and time slots are assumed to be independent.

2) SM–MIMO TMs: As an illustrative example, two specific transmission modes are depicted in Fig. 7.

1) In TM1 (SSK-like transmission [55]), the information bits are conveyed only via the spatial-constellation diagram. Accordingly, \(M = 1\) and, thus, \(A = \text{deg}\). In this case, \(N_{a,\ell}\) for \(\alpha = 1, 2, \ldots, A\) can assume all possible values in the range \(0 \leq N_{a,\ell} \leq N_t\) and, thus, a different number of TAs can be active over different frames. This offers a flexible tradeoff between SE and EE by judiciously choosing the number of RF chains. As mentioned in Section III-C, neglecting the signal-constellation diagram for data modulation reduces the achievable peak throughput, but it can significantly simplify the design and can increase the efficiency of the power amplifiers thanks to the constant-envelope nature of SSK–MIMO communications.

2) In TM2 (SM-like transmission [53]), the information bits are conveyed via both the spatial- and signal-constellation diagrams. In this case, it is convenient to restrict the choice of the spatial-constellation diagram to having elements, \(a^{(\ell)}\), with the same number of nonzero entries, i.e.,
$N_a = N_\alpha$ for $\alpha = 1, 2, \ldots, A$ and $1 \leq N_a \leq N_t$. This assumption is useful for two main reasons: 
a) it allows SM–MIMO to use a single mother space-time code, which is a reasonable choice for simplifying the system design and optimization; and 
b) it allows SM–MIMO to transmit a fixed number of bits per frame, which simplifies demodulation at the receiver. The setup associated with $N_\alpha = 0$ is not considered, since this would lead to decoding ambiguities of the signal-constellation diagram when no TA is active. On the other hand, the setup associated with $N_\alpha = N_t$ is a degenerate case, since it corresponds to $A = 1$ and, thus, data cannot be modulated onto the spatial-constellation diagram. In this case, SM–MIMO in (3) reduces to conventional MIMO communications. This highlights the backward compatibility of SM–MIMO communications: SM–MIMO reduces to standardized MIMO by neglecting the excess (inactive) TA elements. It is important to note that equipping conventional MIMO schemes with the SM functionality comes at a moderate cost since only passive TA elements and a fast RF switch are needed. Thus, SM–MIMO is much less invasive and may be deemed to require less time-to-market than other single-RF [36] and large-scale [63] MIMO solutions.

3) Encoding Algorithm: Given the transmission model in (3), the encoding at the transmitter can be summarized as follows. The information bits are processed in blocks of $\log_2(A) + N_M \log_2(M)$ bits, where $N_M = 0$ for TM1 and $N_M \geq 1$ for TM2. $\log_2(A)$ bits are used to choose a vector $a^{(s)} \in A$ of the spatial-constellation diagram. Furthermore, $N_M \log_2(M)$ bits are used to choose $N_M$ complex symbols, $(\mu_1 \in M, \mu_2 \in M, \ldots, \mu_{N_M} \in M)$, of the signal-constellation diagram. These complex symbols are simultaneously transmitted from the $N_A$ active TAs over $N_t$ time slots according to the mother-space-time code chosen. On the other hand, the symbol $a^{(s)}$ of the spatial-constellation diagram does not change over a frame. The active TAs are kept the same for a frame duration. From one frame to another, the active TAs change according to the incoming $\log_2(A)$ bits and to the spatial-constellation diagram.

Since $N_t$ time slots are used to transmit $\log_2(A) + N_M \log_2(M)$ bits, the system’s rate is $R = (\log_2(A) + N_M \log_2(M))/N_t$ bpcu. In TM2, even though it is clear that OSTBCs have a normalized rate of $N_M/N_t \leq 1$, if the spatial-constellation diagram is a nondegenerate set, i.e., we have $A > 1$, then the normalized rate $R = R/\log_2(M) = \log_2(A)/[N_t \log_2(M)] + (N_M/N_t)$ can be higher than one. Likewise, SM–MIMO is capable of offering rates higher than spatial-multiplexing MIMO with the same number of active RF chains. In fact, if we have $N_M = N_\alpha$ and $N_t = 1$, then $R_{SM} = \log_2(A) + N_\alpha \log_2(M)$ bpcu and $R_{SMX} = N_\alpha \log_2(M)$ bpcu. In summary, high-rate MIMO systems can be designed by appropriately configuring the spatial-constellation diagram and the number of RF sources.

In particular, the proposed spatially modulated space-time-coded generalized SM–MIMO system subsumes several multiple-antenna schemes. For example, assuming $w_1(\xi) = w_0(\xi)$ for $t = 1, 2, \ldots, N_t$, we have:

1) in TM2, if $A = 1$, i.e., $A$ is a degenerate set, then (3) reduces to a conventional space-time-coded MIMO system with a rate of $R = (N_M/N_t) \log_2(M)$ bpcu [51, 52];

2) in TM2, if $A = 1$, i.e., $A$ is a degenerate set, $N_t = 1$, $N_M = N_\alpha$, and $M(\cdot)$ contains $N_M = N_\alpha$ independent symbols, then (3) reduces to a conventional spatial-multiplexing MIMO system associated with independently encoded streams and a rate of $R = N_\alpha \log_2(M)$ bpcu [29, Sec. 2.2.3];

3) in TM2, if $N_t = 1$, $N_M = 1$, and $N_\alpha = 1$ for every $a^{(s)} \in A$, then (3) reduces to SM with a rate of $R = \log_2(N_t) + \log_2(M)$ bpcu [53];

4) in TM1, if $N_t = 1$ and $N_\alpha = 1$ for every $a^{(s)} \in A$, then (1) reduces to SSK modulation with a rate of $R = \log_2(N_t)$ bpcu [55].

Several other design options are also possible, when appropriately choosing the spatial-constellation diagram. These advanced encoding schemes are introduced and discussed in Section IV-B. Finally, it is worth mentioning that an important component of the generalized SM–MIMO in (3) is the adoption of the shaping filters $w_t(\cdot)$ for $t = 1, 2, \ldots, N_t$. The adequate choice of these filters is essential for providing a wide range of multiplexing/diversity gains at a low (single-stream) signal processing decoding complexity. Further information is available in [50, Table 1] and in Section IV-E, where various transmit-diversity schemes designed for SM–MIMO are discussed.

4) ML-Optimum Decoding: Assuming that perfect synchronization and channel state information are available at the receiver, the ML-optimum decoding rule of SM–MIMO formulated in (3) is given in

$$\begin{align*}
(\hat{a}, \hat{\mu}) &= \arg \min_{a^{(s)} \in A, \mu \in \{\mu_1 \in M, \mu_2 \in M, \ldots, \mu_{N_M} \in M\}} \\
&\left\{ A(\hat{a}, \hat{\mu}) = \sum_{s=1}^{N_t} \sum_{r=1}^{N_t} \int_{(s-1)T_t}^{sT_t} \right. \\
&\left. \left|z_{st}(\xi) - \tilde{z}_{st}(\xi)\right|^2 d\xi \right\} \quad (5)
\end{align*}$$

Let

$$\begin{align*}
E_{\tilde{a}} = \sum_{s=1}^{N_t} \left| \sum_{a^{(s)} \in A} \sum_{\mu \in \{\mu_1 \in M, \ldots, \mu_{N_M} \in M\}} \left\| \overline{\Psi}_a \right\|^{2/3} \\
\left| {X}_{st}(\tilde{a}) \overline{H}_r \mu (\xi - (s-1)T_t) \right|^2, \quad &\text{if } \|a^{(s)}\|^2_1 \neq 0 \\
0, \quad &\text{if } \|a^{(s)}\|^2_1 = 0
\end{align*}$$

where

$$\tilde{z}_{st}(\xi) = \frac{\int_{(s-1)T_t}^{sT_t} \left|z_{st}(\xi) - \overline{\Psi}_a \right|^{2/3} d\xi}{\int_{(s-1)T_t}^{sT_t} \left|z_{st}(\xi) - \overline{\Psi}_a \right|^{2/3} d\xi}.$$
where $X^{(\alpha)}(\hat{\mu})$ is the hypothesis at the receiver, which is defined as the $(N_t \times N_r)$-element matrix whose $(s,t)$th entry is $X^{(\alpha)}(\hat{\mu})$; and $(\hat{\alpha}, \hat{\mu})$ is the pair of estimated symbols belonging to the spatial and classic signal-constellation diagrams.

A brute-force implementation of the decoder characterized in (5) needs the computation of $AM^N$ hypotheses (decision metrics $A(\cdot, \cdot)$). Thus, as expected, the complexity of a brute-force search is exponential with the number of information symbolsstreams $N_M$. The interested reader is referred to [50, Sec. IV] and to Section IV-E for some sufficient conditions of choosing spatial-constellation diagram and shaping filters for ML-optimum single-stream demodulation. Furthermore, it is worth mentioning that for those setups for which the SM–MIMO scheme of (3) reduces to conventional MIMOs, e.g., OSTBCs, the decoder formulated in (5) can be simplified and ML-optimum single-stream decoding can be obtained [67].

Finally, we close this section by exploring Figs. 8 and 9, which are reproduced from [50] with permission. These figures compare the proposed generalized SM–MIMO transceiver having a single-stream decoding complexity to many state-of-the-art MIMO schemes relaying, in general,

on multistream decoding complexity. The figures clearly show the potential performance gain of SM–MIMO, despite its low decoding complexity.

E. SM–MIMO and Its Relatives: Single-RF and Large-Scale MIMO Research

As stated in Section II-C, SM–MIMO judiciously combines, with the introduction of the spatial-constellation diagram, the advantages of two promising fields of research in multiantenna communications: single-RF and large-scale MIMO design. In this section, we intend to deepen the connection of SM–MIMO and single-RF/large-scale MIMO research with a twofold objective. 1) On the one hand, we wish to highlight that, in recent years, many solutions have been proposed independently with the aim of simplifying the design of MIMO communications. We believe that this trend reinforces the potential impact of SM–MIMO research in the context of next-generation wireless systems. 2) On the other hand, we wish to harmonize the efforts of independent research for a more fruitful cross fertilization of ideas and results. We believe that this may contribute to a more synergic effort toward the development of a new air interface, which is capable of
offering a flexible tradeoff among EE, SE, error probability, and complexity.

The following MIMO communications paradigms are some notable examples of close relatives of the SM–MIMO principle.

- **BSs with a variable number of TAs.** Limiting the number of active RF chains for the sake of reducing the power consumption of BSs has been actively discussed within 3GPP standardization bodies. In particular, power amplifier switching methods have been proposed in [68]. The main idea is to enable the BSs to use only a subset of the available TAs in order to reduce the power consumption. SM–MIMO shares the same principle, but it still exploits the inactive TA elements either for better SE or for reducing the RF output power.

- **Parasitic single-RF MIMO.** New multiple-antenna designs based on compact parasitic architectures have been proposed for achieving multiplexing gains with the aid of a single active RF source and many passive TA elements. The key idea is to change the radiation pattern of the antenna array at symbol rate and to encode independent information streams onto the angular variations of the far field in the wave-vector domain [40], [69]. Recently, parasitic MIMO systems have been implemented and tested in indoor environments [70], [71]. SM–MIMO shares the single-RF structure of parasitic MIMOs but the information is encoded onto the TA switching mechanism, rather than onto the radiation pattern of the antenna array.

- **Incremental single-RF MIMO.** New MIMO schemes jointly combining multiple-antenna transmission and the ARQ feedback have been proposed to avoid keeping all available antennas active. The main idea is to achieve MIMO gains with the aid of a single-RF chain and a single power amplifier, as well as to reduce encoding and decoding complexity. In particular, in [72], Hesami and Laneman propose a scheme that reduces the system complexity by having a single active antenna at a time, but exploiting the ARQ feedback flag to randomly cycle through the available TAs in case of incorrect data reception. The SM–MIMO philosophy is motivated by the same low-complexity implementation principle, but SM–MIMO has the benefit of being an open-loop scheme, while the incremental single-RF MIMO is a closed-loop scheme using ARQ feedback.

- **ASM.** New directional modulation schemes have been proposed for MMB systems to enable secure and low-complexity wireless communications. The main concept is to develop modulation schemes that randomize the transmitted symbols received by eavesdroppers, while facilitating the reception of the intended users. In particular, in [73], Valliappan et al. introduce an antenna-level modulation technique that eliminates conventional baseband circuitry and takes advantage of the full antenna array with the need for a limited number of RF chains. The idea is to modulate the radiation pattern at symbol rate by driving only a subset of TAs in the array. While randomly switching antenna subsets does not affect the symbol modulation for a desired receiver along the main direction, it effectively randomizes the amplitude and phase of the received symbol for an eavesdropper along a sidelobe. SM–MIMO shares many of the benefits of ASM [73], but it targets higher rates rather than secure communications. Furthermore, ASM exploits directional beamforming to overcome the path loss, the atmospheric absorption, and the high noise levels observed at millimeter-wave frequencies, while state-of-the-art SM–MIMO schemes are not designed to provide beamforming gains. Finally, similar to SM–MIMO, ASM advocates the adoption of constant-envelope transmit signals in order to minimize the linearity requirements of the power amplifiers and to enable high power efficiency by operating near the saturation region.

- **Low-complexity precoding for millimeter-wave communications.** In MMB system design, the cost of implementing one RF chain for every TA can be prohibitive [47]. For this reason, analog baseband beamforming or RF beamforming with one or a few RF chains can be a promising alternative for low-complexity solutions. In particular, in [74] and [75] (and references therein), El Ayach et al. propose low-complexity hybrid RF/baseband precoding schemes where large antenna arrays are driven by a limited number of transmit/receive RF chains. At the transmitter, for example, precoding is split in two parts: first a digital baseband precoding is applied by assuming a limited number of RF chains and, then, a constrained RF precoding is implemented by using simple RF phase shifters. SM–MIMO can be a viable low-complexity modulation scheme for MMB systems for two main reasons: 1) it reduces the number of RF chains compared to conventional MIMOs; and 2) it increases the throughput by capitalizing on the benefits of large antenna arrays that can be packed compactly at both the transmitter and the receiver due to the decrease in wavelength. However, the application of SM–MIMO to millimeter-wave frequencies is not without challenges. In fact, state-of-the-art SM–MIMO schemes provide limited beamforming gains, which, on the other hand, are essential in these frequency bands. Thus, enhanced SM–MIMO schemes, jointly providing multiplexing
and beamforming gains, are needed for application to millimeter-wave frequencies.

- **Large-scale (massive) MIMO.** Large-scale MIMO is an emerging technology, where BSs with hundreds of antennas serve a few mobile terminals per channel use [41]. The main idea behind large-scale MIMO is to reap all the benefits of conventional MIMO, but on a much greater scale [42]. SM belongs to the family of large-scale MIMO systems, since it needs sufficiently large antenna arrays to provide high multiplexing gains. However, SM–MIMO is implemented with far fewer RF chains than the number of available TAs in order to reduce the system complexity and to improve the EE. As a consequence, in SM–MIMO, the number of data streams that can be simultaneously transmitted is smaller than in conventional large-scale MIMOs, since it is limited by the number of RF chains. Accordingly, the peak throughput of large-scale MIMOs is expected to be higher than in SM–MIMO. However, SM–MIMO may trade off this SE suboptimality by avoiding two fundamental problems in large-scale MIMO research [63, Sec. 5]: 1) internal power consumption/dissipation; and 2) economy of scale and low-cost hardware design. In fact, as stated in Section III-C, SM–MIMO reduces both the transmit power and the circuit power consumption, as well as the number of hardware components since only a few RF chains are needed. In that regard, SM–MIMO offers a fundamental SE versus EE/complexity tradeoff, which can be optimized by judiciously choosing the number of active TA elements to meet the desired QoS requirements [76], [77].

IV. PART II—SM–MIMO: A COMPREHENSIVE SURVEY

A. Historical Perspective

Although SM–MIMO has received widespread attention from the research community only in the last four or five years (about 11 papers were available in the open technical literature in 2008), it is a 13-year old technology. In this section, we briefly describe the history of SM–MIMO research with a focus on some pioneering papers, where the concept of SM is first proposed and studied. From this short glimpse into the historical development of SM–MIMO, it is evident that during the years 2001–2008 various researchers independently developed transmission concepts closely related to the SM–MIMO scheme described in Section III-A, which were nicknamed with different acronyms.

The “space modulation” principle appeared for the first time in 2001 [81]. The scheme is called SSK modulation and it exploits the differences in the signals received from different TAs to discriminate the transmitted information messages. In [81], a two-antenna MIMO setup providing 1-bpcu transmission is investigated. The information bits are encoded by keeping one TA active all the time, while activating the second TA only for one of the two possible information bits. The authors also propose a two-antenna MIMO setup that combines spatial and 2-PSK modulations, thus providing 2-bpcu transmission. The scheme employs 2-PSK modulation, when the TAs are active.

A year later in 2002, in [82], Haas et al. proposed a multiantenna modulation scheme, where the number of bits that is equal to that of the TA elements is multiplexed in an orthogonal fashion. A special property of the encoding scheme is that only one out of the available TAs is active in every channel use. The scheme is developed for 2-PSK modulation and some bits are used as parity checks. The authors show that it provides the same spectral efficiency as 8-PSK, but with a better error probability.

Two years later in 2004, in [83], Song et al. proposed a modulation scheme termed as “channel hopping technique,” which is exactly what is known today as SM–MIMO. The idea is portrayed in [83, Fig. 1], and it foresees the transmission of two information streams: the first is explicitly transmitted by using conventional PSK/QAM and the second is implicitly transmitted by activating a single TA of the available antenna array.

In 2005, Mesleh et al. [84] independently proposed the same modulation scheme as in [83]. The main motivation behind [84] is to develop an ICI-free multiantenna modulation scheme, which is realized by activating one TA in every channel use and by encoding some information bits using the TA switching process. The authors also show that the proposed scheme significantly relaxes the signal processing complexity at the receiver.

In 2006, the same authors further investigated the scheme proposed in [84], and they used for the first time the terminology of “spatial modulation” to identify this encoding mechanism [85]–[88].

Two years later in 2008, various papers were published by improving and further investigating the SM–MIMO concepts presented in the previous years. In [89], the authors move from [83] and study the channel capacity. In this paper, the parlance of IGCH is coined. It is shown that IGCH provides better spectral efficiencies than OSTBCs. In [53], the SM–MIMO concept introduced in [84]–[88] is comprehensively studied by using a low-complexity two-step demodulator. This paper shows the potential advantages of SM–MIMO compared to state-of-the-art spatial-multiplexing and Alamouti schemes. In [56], Jeganathan et al. develop the ML-optimum demodulator for SM–MIMO, and they show that some performance improvements can be expected compared to the suboptimal demodulator introduced in [53]. In [55], the SM–MIMO scheme is simplified by generalizing the SSK concept originally proposed in [81] to arbitrary numbers of TAs. In particular, only the spatial-constellation diagram is
used to transmit information bits. This encoding scheme is today known as SSK–MIMO. In [90], a framework is introduced for the performance analysis of SM–MIMO by using results from ordered statistics. In [91], Jeganathan et al. extend [55] by allowing more than one TA to be active in every channel use and by encoding the information bits onto various combinations of multiple active TAs. They show that for the same number of TA elements the rate can be improved at the cost of increasing the number of RF chains, while tolerating some performance loss. This modulation scheme is referred to as GSSK.

We conclude this glimpse into the history of SM–MIMO by mentioning that the first (and only available) survey papers on the SM concept were published in December 2011 and in mid 2012 in [48] and [92], respectively. On the other hand, the world’s first measurement campaign and experimental testbed substantiating the benefits of the SM principle in realistic outdoor and indoor propagation environments appeared in May 2013 in [93] and [94], respectively.

**B. Transmitter Design**

At the time of writing, more and more modulation schemes emerge with a focus on exploiting either a single TA index or combinations of TA indices to convey data. All these schemes are generalizations of the solutions in [53], [55], and [91] and discussed in Section IV-A. They strike, in general, a tradeoff between the achievable rate, the system performance, the encoding/decoding complexity, as well as the number of active RF chains and, thus, the overall EE. For example, let us again denote by $N_t$ the number of TA elements at the transmitter, by $N_{RF}$ the number of available RF chains, and by $M$ the signal modulation order. Recall from Section III that the rate is $R = \log_2(N_t) + \log_2(M)$ bpcu for SM–MIMO and $R = \log_2(N_t)$ bpcu for SSK–MIMO, respectively. Furthermore, in both cases, we have $N_{RF} = 1$. On the other hand, the GSSK–MIMO scheme proposed in [91] encodes the information bits onto combinations of multiple active TAs. This hence results in a rate equal to $R = \lfloor \log_2(N_{\text{comb}}) \rfloor$ bpcu, where $\lfloor . \rfloor$ is the floor-function and $N_{\text{comb}} = \left( \frac{N_t}{N_{RF}} \right)$ with ( ) denoting the binomial coefficient. Thus, at the cost of increasing the number of RF chains, GSSK–MIMO provides higher rates than SSK–MIMO. However, this encoding scheme still preserves the ICI-free advantage even though more than one TAs are active. In general, the choice of the spatial-constellation diagram is not unique in GSSK–MIMO, and it can be optimized in order to minimize the error probability [91]. Inspired by [53], [55], and [91], various encoding mechanisms, link-adaptation schemes, and transmit-preprocessing solutions have been proposed [95]–[142].

In [102] and [104], the authors combine the SM–MIMO and GSSK–MIMO concepts to further improve the achievable spectral efficiency. The proposed modulation scheme is termed as GSSM–MIMO. The main idea is to modulate the information bits onto both the signal-constellation diagram and combinations of multiple active TAs. The price to pay is a higher decoding complexity, but the ICI-free transmission is still preserved.

In [114] and [141], the authors propose some encoding schemes, where the number of active TAs can be different in different channel uses. The modulation proposed in [114] is SSK-like, hence no signal modulation is used. On the other hand, both the signal- and spatial-constellation diagrams are used in [141]. However, the ICI is avoided in both cases, since a single point of the signal-constellation diagram is transmitted in every channel use. By allowing a variable number of TAs to be active, the rate can be significantly increased. For example, by assuming that the number of active TAs can span the whole range $0 \leq N_{RF} \leq N_t$, the SSK-like modulation schemes of [114] and [141] can offer a rate equal to $R = N_t$ bpcu, which shows that the rate no longer increases logarithmically but linearly with $N_t$. By adding the signal-constellation diagram [141], higher rates can be obtained. The main drawback of these schemes is that, in order to achieve these peak rates, the transmitter must be equipped with a number of RF chains that is equal to $N_{RF} = N_t$. This implies that a more modest reduction of the static power consumption is expected [30]. Furthermore, the encoding and decoding complexities are increased. In [118], SM–MIMO is combined with spatial multiplexing for further increasing the rate at the cost of increasing the decoding complexity. The main idea is to allow the transmission of multiple symbols of the signal-constellation diagram during the same channel use, and, at the same time, to encode some bits onto the indices of the active TAs. The main advantage of this transmission scheme is its increased rate achieved at the cost of adding ICI. In [138], it has been recently shown that, given $N_t$, an optimal number of RF chains may exist that maximize the rate, while obeying $N_{RF} < N_t$.

The encoding schemes of [53] and [55] require that the number of TAs is a power of two. This restriction can be avoided by using the generalized SSK/SM-like modulation schemes of [91] and [102]. However, the price to pay is the need for multiple active TAs, and, thus, RF chains. In [98], this problem is avoided by introducing a non-integer-based encoding mechanism, which relies on the application of modulus conversion to achieve fractional bit rates. The main drawback of this solution is the presence of an error propagation effect. This issue is solved in [105] by proposing a bit padding method. The idea is to map different bit lengths onto the indices of the TAs and then to use padding techniques for avoiding error propagation. This approach can be used for arbitrary numbers of TAs without error propagation.

As mentioned in Section III-C, SM–MIMO requires the impulse responses of the transmit-to-receive links to be
sufficiently different from each other for the sake of low demodulation error probability. For example, in the presence of spatial correlation, the channel impulse responses may be too similar to each other to guarantee a good detection performance. In [100] and [111], an encoding mechanism based on TCM is proposed for reducing the impact of channel correlation on the system’s performance. The proposed scheme is termed as TCSM. In particular, TCM is used in conjunction with SM to partition the TAs into subsets by maximizing the spacing between TAs of the same subset. In this approach, only the specific information bits that determine the TA index are convolutionally encoded. The main advantage of this encoding is that it offers better performance in correlated channels. However, it does not provide any error probability advantage in uncorrelated channel conditions, compared to uncoded SM–MIMO. This limitation is circumvented in [111], where Basar et al. propose a different TCM-based method, which benefits from the advantages of trellis coding in both uncorrelated and correlated fading channels.

Recently, various encoding methods have been proposed that aim for improving the performance of SM–MIMO by exploiting CSIT. In [96], Handte et al. generalize SM–MIMO by proposing an encoding scheme, where a beamforming vector is chosen from a given codebook based on the bits to be transmitted, rather than a single TA. The optimal precoding matrix is chosen by minimizing the average error probability, which depends on the CSIT. The authors show a promising performance improvement, which is achieved at the cost of activating multiple TAs. In [99], Di Renzo and Haas solve the optimal power allocation problem for SSK–MIMO with two TAs. They show that if the average CSIT knowledge is available, the optimal transmission scheme results in an on–off encoding, where the transmit power should be chosen as a function of the power imbalance ratio and correlation coefficient of the wireless links. In [140], the system model of [96] and [99] is generalized and the optimal constellation design associated with instantaneous CSIT knowledge is solved for general MIMO setups. The authors consider both perfect and imperfect CSIT, and solutions with one and many active TAs are studied. Substantial performance gains are obtained compared to the CSIT-unaware scenario. In [110] and [128], link-adaptive solutions are introduced and studied. In particular, in [128], Yang et al. study transmission modes for SM–MIMO communications where, given a specific constraint on the desired rate, various modulation orders and number of active TAs can be chosen. Diverse solutions are proposed in order to trade off the feedback overhead against the performance gain. Overall, the authors show a significant improvement in the system performance.

To further improve the performance of SM–MIMO, various transmit-preprocessing schemes are proposed in [107], [119], and [121]. In contrast to the SM–MIMO and SSK–MIMO schemes, which carry information using the indices of the TAs and assume the availability of channel state information at the receiver (CSIR), these papers propose to encode the information bits onto the indices of the RAs and assume CSIT. To this end, precoding/beamforming matrices are designed by using CSIT in order to focus all the transmit power onto a single RA in every channel use. The first paper proposing the idea is [107], and Yang shows that the proposed scheme may provide a better performance than open-loop schemes. In [119], the proposal of [107] is further generalized by assuming imperfect CSIT and two new precoders are proposed. In [121], the scheme is studied in the presence of frequency-selective fading channels, and it is shown that the larger the number of TAs the better the performance. The main limitation of these schemes is the need for a large number of TAs and RF chains compared to the small number of RAs. In fact, the multiplexing gain depends on the number of RAs.

Inspired by the SM–MIMO principle, Sugiura et al. [103], [106] propose an encoding scheme based on a dispersion matrix approach. In particular, rather than mapping the information bits onto TA indices, the dispersion matrix approach of [103] spreads the bits with the aid of space-time-domain matrices, which are appropriately optimized to provide the desired performance. The authors show that the dispersion matrices can be chosen in order to satisfy a specific ICI constraint at the receiver to facilitate ML-optimum single-stream detection. Furthermore, they show that either multiple TAs or a single active TA can be accommodated by trading off performance against complexity. If the transmission spans more than one time slot, the authors show that transmit-diversity gains can also be achieved. The adoption of the dispersion matrix approach requires the appropriate design of spreading matrices with specific characteristics in order to provide the desired performance. In [103], a Monte-Carlo-simulation-based random search algorithm is employed for the sake of simplicity. However, due to the large number of parameters to be optimized, the computational complexity can be quite high. For this reason, more recently, various authors have proposed more efficient approaches for the construction of such matrices [143]–[149]. Specifically, in [143], the coding gain is used as the metric to be maximized, while maintaining the diversity order. In [145], genetic algorithms are proposed in order to efficiently guide the random search algorithm through the associated large design space. In [147], the dispersion matrices are jointly designed with the signal constellation, by allowing a substantial reduction of the computational complexity. In [149], a systematic approach is proposed for the design of the dispersion matrices, which does not require any numerical search and it is applicable to MIMO systems with more than three TAs. The dispersion matrices provide a second-order transmit diversity.
C. Receiver Design

The demodulation algorithm originally proposed in [53] for estimating both the active TA index and the transmitted symbol from the signal-constellation diagram is based on a two-step approach: 1) first, the TA-index is estimated; and 2) then, based on the estimated active TA, the signal-constellation symbol is demodulated. This approach offers low-complexity demodulation. However, it is suboptimum and it is affected by a high sensitivity to demodulating the TA index. To overcome these limitations, in [56], Jeganathan et al. propose the ML-optimum demodulator that jointly decodes both the TA index and the transmitted symbol. Recently, various demodulation schemes have been proposed with the main objective of attaining near ML-optimum performance at a low demodulation complexity [150]–[170].

In [155], Younis et al. advocate low-complexity demodulation schemes based on the SD principle. Two SDs, namely the Rx-SD and Tx-SD, are proposed, which aim for reducing the search space of the number of RAs and TAs, respectively. Via a proper choice of the parameters, it is shown that both demodulators provide a substantial reduction of the computational complexity, while retaining the same performance as ML-optimum decoding. Simulation results show that no SD is superior to each other and for every MIMO setup, but the best SD to choose depends on the number of TAs and RAs, as well as on the received SNR. The rule of thumb is that the Rx-SD is more suitable for low spectral efficiencies, while the Tx-SD is the best choice for high spectral efficiencies. The SDs are useful for arbitrary numbers of TAs and RAs. Similar SDs have been recently proposed in [159] for SM–MIMO.

In [162], Rajashekar and Hari propose an ML-optimum demodulator for SM–MIMO, whose complexity is independent of the size of the signal-constellation diagram. The solution is applicable to either square or rectangular lattice constellations, such as QAM. The idea is to perform hard limiting directly on the received signal without the need for searching through the signal-constellation diagram. The demodulator is applicable to general MIMO setups associated with a single active TA. Furthermore, the authors combine the proposed hard-limiting demodulator with SD in order to further reduce the complexity of demodulating the spatial-constellation diagram.

In [165], Yu et al. conceive a low-complexity demodulator for SSK–MIMO systems based on the CS principle [171], [172]. The idea is especially suitable for large-scale MIMO implementations with a very large number of TAs, while only a few of them are active in each channel use. The idea originates from the observation that for these MIMO setups the transmission vector of SSK–MIMO is sparse, i.e., there are many zero entries that correspond to the inactive TA elements. Thus, sparse signal processing methods based on CS can be exploited. For example, demodulation can be performed by minimizing the 1-norm and by applying the OMP algorithm [173]. The authors show that the performance can be significantly reduced at the cost of some performance degradation. The key parameter to be carefully chosen in order to strike a flexible tradeoff between performance and complexity is the number of RAs: the larger the number of RAs, the better the performance.

The two-step MF-based and ML-optimum demodulators introduced in [53] and [56] cater for two extreme situations: the first demodulator offers a low computational complexity, while the second offers the best performance. Recently, various demodulators have been proposed for striking the right tradeoff between these two extremes. In [154], Sugiura et al. propose two improved versions of the MF-based demodulator of [53]: the EMF and NMF detectors. Both detectors provide better performance than the conventional MF detector. In particular, the EMF demodulator offers a better performance than the NMF demodulator at the cost of increasing the signal processing complexity. More recently, in [161], Yang et al. further improve the solutions proposed in [154] by carefully choosing a design parameter, i.e., the number of most probable TA indices that can trade off performance against signal processing complexity.

The reason behind the performance difference between the two-step and the ML-optimum demodulation in [53] and [56], respectively, originates from the fact that even though the TA-index and PSK/QAM symbol are encoded independently in SM–MIMO, they fade together during transmission through the channel. As a consequence, any attempt of demodulating them independently results in a nonnegligible performance loss. To overcome this limitation, while dispensing the receiver with the signal processing complexity of ML-optimum demodulation, in [169], Xu et al. develop a decoding algorithm that allows separate decoding of the spatial- and signal-constellation diagrams, while taking into account their correlation. The authors propose both hard- and soft-decision solutions, and they show that the optimal performance can still be retained, despite its low demodulation complexity.

D. Error Performance and Capacity Analysis Over Fading Channels

Being a new modulation concept conceived for MIMO communications, the performance of SM–MIMO has been studied extensively in the last few years [53], [55], [56], [64], [66], [90], [96], [174]–[193]. The main goal has been fourfold: 1) to understand the impact of wireless propagation on the end-to-end error probability and achievable rate; 2) to identify those propagation scenarios, where the adoption of SM–MIMO is a suitable choice; 3) to provide guidelines for new system designs and transmission concepts based on the SM–MIMO principle; and 4) to shed light on the fundamental properties such as the achievable diversity order.
The first attempts to study the performance of SM–
MIMO were reported in [90], which were later general-
ized and extended in [53]. Therein, the authors use
ordered statistics to propose a framework that computes
the error probability of the two-step demodulator of [53]
for transmission over Rayleigh fading channels. The
framework is semianalytical and requires the calculation
of some integrals. This framework is generalized in [175]
for transmission over Nakagami-m fading channels.

In [55] and [56], Jeganathan et al. propose a framework
for studying the performance of the ML-optimum demod-
ulator for transmission over Rayleigh fading channels. In
particular, the framework of [55] is applicable to SM-like
modulation schemes, while the framework of [56] is appli-
cable to SM-like schemes associated with a real-valued
signal-constellation diagram. These frameworks highlight
that the error probability of SM–MIMO depends on the
Euclidean distance of pairs of channel impulse responses,
as well as that SM–MIMO may exploit the receive
diversity better than conventional single-antenna com-
munications. Furthermore, the diversity order is shown
to be equal to the number of RAs regardless of the number
of TAs.

In [96], Handte et al. propose a general framework for
studying the performance of SM–MIMO for transmission
over correlated Rician fading channels. They use random
matrix theory for computing the pairwise error probability
and then exploit the union-bound-based method for
evaluating the error probability. The framework is
applicable to general signal-constellation diagrams. The
main insight of this framework is that SM–MIMO is
unable to achieve any transmit-diversity gain even though
multiple TAs are simultaneously activated. A similar result
is obtained in [178] for various SM–MIMO schemes.
Therein, some initial solutions are provided for achieving
transmit diversity. These results motivated many research-
ers to propose improved SM–MIMO schemes providing
both receive and transmit diversities, as described in
Section IV–E.

In [55], [100], and [177], the authors study the per-
formance of channel-coded SM–MIMO communications.
In particular, in [55], BICM-aided schemes are studied.
The authors show that the achievable diversity is still
independent of the number of TAs, but it depends on the
number of RAs and on the Hamming weight of the code. In
[100] and [177], the authors study the performance of
TCSM using hard-decision Viterbi decoding. The mathe-
atical framework is applicable to convolutional coding,
and it exploits the augmented transfer function method for
performance analysis. As such, it is applicable to general
convolutional codes. The analysis reveals that the diversity
order of TCSM depends on the free distance of the con-
volutional code adopted.

In [64], [66], and [182], a comprehensive mathe-
matical framework is introduced to study the performance
of SSK–MIMO communications for transmission over
Nakagami-m and Rician correlated fading channels. The
fundamental performance trends of SSK–MIMO systems
are unveiled. More specifically, the following performance
trends are derived analytically. 1) The error probability
improves if there is power imbalance among the TAs. This
result originates from the fact that the error probability
depends on the difference between pairs of channel im-
pulse responses. This outcome has been recently exploited
for developing optimal power allocation schemes for per-
formance improvement [99], [140]. 2) The error proba-
bility over identically distributed fading channels in the
absence of power imbalance typically degrades if the
channel correlation increases. This result is expected since
the channel impulse responses are similar to each other
and hence demodulation becomes less reliable. 3) On the
other hand, for transmission over nonidentically distrib-
uted fading channels subject to power imbalance, the error
probability may improve in the presence of channel
correlation. More specifically, the trend depends on how
deep the channel fluctuations are. In general, the less deep
is the fading, the better is the performance. The intuitive
reason behind this trend is that power imbalance makes
the channel impulse responses different from each other
and that random channel fluctuations reduce 
the Euclidean distance. 4) Over Rician fading channels, the
performance of SSK–MIMO degrades when increasing the
Rician factor. This behavior is in contrast to conventional
modulation schemes transmitting over Rician fading,
where the higher the Rician factor the better the perform-
ance. The reason behind this trend is that the higher the
Rician factor, the less different the channel impulse
responses, thus the worse the error probability. 5) Over
Nakagami-m fading channels, the diversity order is
independent of the fading severity \( m \). Once again, this
result is different from the trends in conventional modu-
lation, where the diversity order linearly increases with \( m \).
These results clearly show that some channels are more
suitable than others for SSK–MIMO communications, and
that improved schemes may be needed for counteracting
the impact of adverse fading conditions.

In [57], the study of [64], [66], and [182] is extended to
SM–MIMO communications over generalized fading
channels. The proposed framework is conveniently formu-
lated as the summation of three terms: the first depending
on the signal-constellation diagram only; the second de-
pending on the spatial-constellation diagram only; and the
third being a function of both constellation diagrams. This
desired-form analytical formulation unveils the following
fundamental performance trends: 1) SM–MIMO has the
inherent potential of providing a better performance than
single-antenna communications if the required throughput
is greater than 2 bpcu and at least two antennas are
available at the receiver; 2) the larger the number of TAs,
the higher the gain compared to single-antenna commu-
nications; 3) the diversity order for transmission over
Rician fading is determined by the number of RAs \( N_r \),
while it is equal to \( \min \{ N_r, m N_t \} \) for transmission over Nakagami-\( m \) fading; this result shows that for fading channels that are less severe than Rayleigh fading \( (m > 1) \), SM–MIMO offers a more modest diversity order than single-antenna systems; on the other hand, it provides a higher diversity order and thus a better performance for more severe fading channels \((0.5 \leq m < 1)\); and 4) the adoption of PSK modulation for the signal-constellation diagram may provide a better performance than QAM for diverse MIMO setups. This outcome is in contrast to conventional modulation schemes, where QAM is always superior to PSK modulation. This finding is important because it brings to our attention that constant-envelope modulation may be a suitable choice for SM–MIMO communications since it may offer an improved power efficiency and a reduced complexity, while mitigating the linearity requirements of the power amplifiers [5], without any performance degradation compared to amplitude modulation schemes. The beneficial impact of using constant-envelope modulation is a topic of current research interest in the context of large-scale MIMO design aiming at reducing the implementation complexity and at improving the EE [194], [195]. Overall, the main outcome of [57] is that, for a given throughput in bpcu, the optimal size of signal- and spatial-constellation diagrams should be chosen as a function of the channel characteristics. For example, if \( m \gg 1 \), more bits should be modulated on the signal-constellation diagram in order to exploit the benign fading of the channel. On the other hand, if \( m \ll 1 \), more bits should be modulated onto the spatial-constellation diagram in order to improve both the coding and diversity gains. Some results concerning this optimization and adaptive transceiver design are available in [132].

In addition to the error probability, many authors have focused their attention on investigating the information-theoretic limits of SM–MIMO communications, by considering both data and secrecy capacity performance metrics. In [190], Yang and Jiao compute the channel capacity of SM–MIMO for Gaussian input signals. The capacity is expressed as the summation of two terms, which account for both signal- and spatial-constellation diagrams. The authors highlight that the channel capacity is almost the same as that of the transmission concept known as “spatial cycling using one transmitter at a time” [54, p. 317]. Accordingly, they show that the information-theoretic capacity is higher than that of single-antenna communications and STBCs equipped with more than two TAs. In [191] and [192], Sinanovic et al. focus their attention to the secrecy rate of SM–MIMO systems in the presence of an eavesdropper. The analysis in [191] shows that the outage secrecy capacity of SM can be better than that of its single-antenna counterpart. In particular, the gains increase upon increasing the number of TAs. Furthermore, in [192], the effect of the constellation size is investigated. It is shown that its effect depends on the values of both the desired user’s and on the eavesdropper’s SNRs. For a low eavesdropper SNR, smaller constellations perform better than larger ones for most of the SNR range, while for a high eavesdropper SNR, larger constellations provide larger secrecy capacities. Finally, evolving from [107], in [193], advanced transmit-preprocessing methods are conceived for physical-layer security based on SM–MIMO transmission. The authors develop three approaches for designing preprocessing weights, including channel diagonalization, eigenvalue-based decomposition, and generalized SM. It is shown that they can strike different tradeoffs between the achievable performance and security. Overall, these results show that SM–MIMO is indeed capable of improving both the spectral efficiency and the secrecy rate, thanks to its unique transmission principle.

E. Transmit-Diversity and Space-Time-Coded Transmission

As mentioned in Section IV-D, the performance analysis of SM–MIMO for transmission over various fading channel models has revealed that no transmit-diversity gains can be expected [96]. This result has motivated researchers to investigate new encoding mechanisms that are capable of improving the performance with the aid of transmit diversity, while still guaranteeing low-complexity single-stream demodulation [50], [64], [80], [103], [106], [178], [196]–[207].

In [80], Basar et al. combine the SM–MIMO concept with the Alamouti code [51] for the sake of proposing a modulation scheme that can take advantage of the multiplexing gain of SM and the second-order diversity of the Alamouti code. The proposal advocates the transmission of an Alamouti code from pairs of active TAs chosen from a spatial-constellation diagram. The second-order diversity is guaranteed by the appropriate optimization of the spatial-constellation diagram. To this end, some rotation angles are introduced and optimized for ensuring both the maximum diversity order and coding gain. The authors show that the proposed SM-aided STBC scheme offers a normalized rate higher than one and can be decoded at a single-stream complexity, while still guaranteeing ML-optimum performance. Numerical results show that it provides a better performance than other state-of-the-art space-time-coded MIMO schemes at the cost of a larger number of TAs without increasing the number of RF chains.

In [103] and [106], Sugiuira et al. introduce the dispersion matrix approach described in Section IV-B. The rate of this scheme is not limited by the number of physical TA elements but by the number of available dispersion matrices. The authors prove that the achievable diversity is equal to \( N_r \min \{ N_s, N_t \} \), where \( N_t \) is the number of TAs, \( N_s \) is the number of RA elements, and \( N_e \) is the number of time slots. This results in a transmit-diversity order equal to \( \min \{ N_s, N_t \} \), which implies that increasing \( N_t \) beyond \( N_s \) does not result in any further transmit-diversity improvement. By contrast, a lower \( N_t \) value may have the twofold
merit of a low computational complexity and of a high transmission rate.

In [198], transmit-diversity schemes are conceived for SSK–MIMO communications with a single active TA per channel use. The encoding scheme is configured for transmission over \(N_t\) time slots. If the number of TAs is higher than two \((N_t > 2)\), the authors show that the achievable transmit-diversity order cannot be higher than \(N_t - 1\). Furthermore, they provide necessary and sufficient conditions for achieving a transmit-diversity order equal to \(N_t\). Specific bit-to-antenna mapping functions are constructed for providing transmit-diversity order equal to \(N_s\), if \(N_s \leq N_t - 1\). If \(N_t = 2\), the authors propose a closed-loop scheme based on phase rotations that provides second-order transmit diversity. This closed-loop scheme has recently been extended in [137] and [199] to more than two TAs. It is shown in [137] that for some rates and antenna setups, full transmit diversity can be achieved by using cophasing and phase rotation. However, accurate CSIT is required.

In [50], [64], [178], [196], and [197], the authors propose a transmit-diversity scheme for SSK–MIMO and SM–MIMO communications that exploits the concept of time-orthogonal shaping filters, and they show that transmit diversity can be obtained without reducing the rate. More specifically, it is shown in [64] that the SSK–MIMO relying on time-orthogonal shaping filters, henceforth denoted by TOSD–SSK–MIMO, is capable of providing second-order transmit diversity, while using a single time-slot transmission and a single active TA element. The authors show that while the error probability of SSK–MIMO degrades upon increasing the Ricean factor, the error probability of TOSD–SSK–MIMO significantly improves. This makes TOSD–SSK–MIMO a more suitable transmission scheme for line-of-sight scenarios. The TOSD principle is generalized in [50], [196], and [197] for achieving transmit-diversity gains higher than two and for application to SM–MIMO communications. In particular, in [50], the generalized SM–MIMO transceiver introduced in Section III-D is detailed and it is analytically shown that a wide range of transmit-diversity and multiplexing gains can be obtained, while still requiring single-stream decoding complexity and still guaranteeing ML-optimum performance. More specifically, two main results are proven: 1) if the information bits are conveyed by only the spatial-constellation diagram (SSK–MIMO), a transmit-diversity order equal to twice the number of active TA elements can be obtained without any rate reduction by using time-orthogonal shaping filters across all the TAs and by constructing the spatial-constellation diagram as a partition, i.e., the points of the spatial-constellation diagram have no TA-elements in common; and 2) if the information bits are conveyed by both spatial- and signal-constellation diagrams, SM–MIMO relying on time-orthogonal shaping filters and a partitioned spatial-constellation diagram may be combined with OSTBCs for the sake of achieving the same transmit diversity as the mother OSTBC, while offering a higher normalized rate. Also, the adoption of time-orthogonal shaping filters facilitates ML-optimum single-stream decoding performance at a low complexity. Furthermore, a simulation comparison with many state-of-the-art space-time-coded MIMO systems is conducted, and it is shown that the SM–MIMO scheme provides an improved performance with the same number of RF chains and lower decoding complexity.

In [203], an open-loop transmit-diversity scheme is designed for SM–MIMO communications that has the beneficial property of providing transmit-diversity gains with the aid of a single active TA element. The proposal is based on the so-called CIOD, which refers to a class of symbol-by-symbol decodable codes that offer full rate (one complex symbol per channel use) for MIMO systems having up to four TAs. The authors combine SM with CIOD and prove that the proposed scheme guarantees second-order diversity at the same rate as SM–MIMO. This is achieved by using a two-slot transmission scheme and by introducing phase rotations for the sake of guaranteeing full diversity. Furthermore, the authors propose low-complexity decoding schemes.

In [206] and [207], the authors combine SM–MIMO with TA selection in order to exploit the advantages of both technologies, resulting in both spatial multiplexing and transmit diversity, respectively. The authors study two criteria for TA selection based on the Euclidean distance and on capacity-based performance metrics. It is shown that capacity-based TA selection provides a better performance and a higher transmit-diversity order.

F. Analysis and Design With Imperfect Channel Knowledge at the Receiver

The vast majority of demodulators presented in Section IV-C rely on the simplifying assumption of having perfect CSIR for their correct operation. However, CSIR is usually affected by CE errors, thus the demodulation performance depends on its accuracy, which is determined by the number of pilot symbols that can be allocated for CE. In addition, due to the encoding mechanism that maps information bits onto TA indices, SM–MIMO may be less robust to CE errors than conventional modulation. To shed light on these fundamental issues, many researchers have investigated the performance of SM–MIMO in the presence of CE errors [55], [62], [208]–[215]. The important conclusion from all these independent research activities is that SM–MIMO may in fact be more robust to CE errors than conventional modulation. This result is quite relevant since it contradicts the basic intuition concerning the effects of CE errors on the performance of SM–MIMO communications.

In [208], Ulla Faiz et al. study the performance of SM–MIMO by simulation with the aid of recursive least squares adaptive CE and compare it to spatial-multiplexing MIMO.
systems. The results show that a gain between 1.5 and 3 dB can be expected.

In [212], Basar et al. corroborate the results obtained in [208] with the aid of an analytical framework. In particular, two receivers are studied in the presence of Gaussian CE errors: 1) the optimal demodulator that exploits the knowledge of the correlation coefficient between the true and estimated channels; and 2) the mismatched demodulator that uses the same decision metric as for perfect CSIR. Both the analysis and the simulations show that SM–MIMO is more robust than spatial-multiplexing MIMOs for practical values of CE errors. The authors also show that the impact of CSIR depends on whether the CE error is dependent or independent of the SNR at the receiver, which actually depends on the specific choice of the channel estimator.

In [213], Ikki and Mesleh study the performance of SSK–MIMO over Rayleigh fading channels in the presence of Gaussian CE errors at the receiver. A closed-form expression of the error probability is provided for arbitrary MIMO configurations. Based on their analytical framework, the authors draw conclusions similar to these of [212]: the diversity order of SSK–MIMO relying on multiple RAs is preserved, if the CE error decreases with the operating SNR. Otherwise, an error floor emerges.

In [62], Di Renzo et al. develop a general analytical framework for the performance of SSK–MIMO and TOSD–SSK–MIMO schemes for transmission over general fading channels and in the presence of a pilot-based ML-optimum channel estimator. The receiver is assumed to be equipped with a mismatched detector, thus resulting in a worst case scenario. The analytical framework clearly shows that SSK–MIMO has the same robustness to CE errors as single-antenna systems. On the other hand, it is shown that TOSD–SSK–MIMO is more robust to CE errors than the Alamouti scheme. Furthermore, it is pointed out that only a low training overhead is required for achieving sufficiently reliable channel estimates for data detection, and that the transmit- and receive-diversity orders of the SSK–MIMO and TOSD–SSK–MIMO schemes are preserved even with imperfect channel knowledge. This conclusion is conditioned on using an ML-optimum channel estimator, which is unbiased and whose CE error variance monotonically decreases with both the operating SNR and the number of pilot symbols transmitted.

G. Performance and Design for Dispersive Channels

In the previous sections, we have provided a comprehensive overview of the research activities dedicated to the analysis and design of SM–MIMO communications over a wide variety of fading channels. Furthermore, diverse demodulation options have been presented. As far as the channel is concerned, the research results presented so far are based on a frequency-flat or narrowband channels. However, in many practical applications conceived for future cellular communications, the wireless channel may be highly frequency selective. This results in a dispersive-channel-induced ISI that may significantly deteriorate the system error probability, if the transceivers are not appropriately designed. Motivated by this consideration, some researchers have recently proposed SM–MIMO transceivers that are robust to frequency-selective fading channels [216]–[221].

In [216], a pair of modulation schemes referred to as SFSK and STFSK modulations are conceived, which represent a generalization of the STSK modulation scheme relying on the dispersion matrix approach of [103]. The main idea behind these transmission schemes is to spread the transmitted signal not only across the space-time domain, but also across the frequency domain. This has the additional benefit that more bits can be encoded onto the space-time-frequency domain, and thus the achievable throughput may be increased. Furthermore, provided that the channel fading in the space-time-frequency domain is independent, the diversity is increased. The authors show that while the performance of STSK modulation is degraded by about 2 dB, when the channel changes from uncorrelated frequency-flat fading to frequency-selective fading, the proposed SFSK and STFSK schemes are capable of maintaining their performance also in frequency-selective fading environments. It is also shown that the STSK and SFSK schemes are special cases of STFSK modulation.

In [217], Ngo and Hanzo then combine the STFSK modulation concept introduced in [216] with SFHMA in order to improve the attainable system performance by providing frequency diversity and by beneficially randomizing the impact of cochannel interference. The ASE of the proposed transmission concept is studied by considering various frequency reuse structures adopted in cellular networks. The ASE of STFSK is then compared against that of state-of-the-art multiple-access techniques often used in cellular networks, including GMSK-aided SFHMA, GMSK-assisted TDMA/FDMA, 3G WCDMA, and 4G LTE. The results show that the ASE of the STFSK-aided SFHMA system is higher than that of GMSK-aided SFHMA, TDMA/FDMA, and WCDMA systems, albeit it is only 60% of the ASE of LTE systems.

It is demonstrated in [218] that the performance of STSK modulation typically degrades in dispersive channels, and the authors propose an improved scheme that jointly combines OFDM and STSK for the sake of avoiding this limitation. The authors show with the aid of Monte Carlo simulations that the proposed OFDM-aided STSK scheme offers a similar performance as single-carrier STSK modulation for transmission over low-mobility narrowband channels. A good performance is observed also for transmission over correlated time-variant scenarios, provided that the parameters are appropriately selected. The authors also discuss the generalization of the proposed scheme to STFSK modulation.
In [219], Hanzo et al. propose a novel OFDMA SC-FDMA-aided multiuser STSK–MIMO architecture for frequency-selective wireless channels, which strikes a flexible diversity versus multiplexing gain tradeoff. The proposed scheme performs well in dispersive channels, while supporting multiple users in a multiple-antenna-aided wireless system. The philosophy of SC–FDMA is to circumvent the problem of high PAPR with the aid of single-carrier transmission and hence use relatively power-efficient class-AB power amplifiers instead of a class-A power amplifier. Furthermore, SC–FDMA facilitates single-tap multiplicative channel equalization because each subcarrier of the parallel modem experiences a non-dispersive narrowband channel. Thus, the overall STSK-based MIMO scheme exhibits a performance similar to that in case of narrowband channels, despite operating in a wideband scenario. The authors also introduce an appropriate mapping of the users’ symbols to the subcarriers in order to support multiuser communications, while requiring low-complexity single-stream-based demodulation.

In [221], the performance of single-carrier SM–MIMO communications is investigated for transmission over frequency-selective fading channels, and it is compared to that of the classic cyclic-prefix and zero-padding based schemes. It is shown that zero-padded SM–MIMO is capable of achieving the maximum attainable multipath diversity provided that ML-optimum demodulation is used at the receiver. On the other hand, cyclic-prefix-based SM–MIMO is not able to provide full diversity. Furthermore, it is shown that the complexity of zero-padded SM–MIMO is independent of the frame length and depends only on the number of multipath echoes.

H. Performance and Design for Multi-User Scenarios

In Section IV-B–F, many SM–MIMO schemes have been studied by implicitly assuming a single-user transmission. However, this operating scenario is quite restrictive for typical cellular deployments, where many users may simultaneously transmit over the same resource block, aiming at maximizing the aggregate throughput at the cost of increasing the interference. Motivated by this consideration, the performance of both optimal and suboptimal receivers designed for SM–MIMO communications has been investigated in the presence of multiple-access interference [222]–[226]. The fundamental issue that these papers intend to address is whether both data modulation and multiple access are possible by relying only on the differences of the TA/RA channel impulse responses and to which extent this is determined by the receiver design.

Di Renzo and Haas [223] study the error probability of SS–K MIMO by considering two receivers: 1) the single-user receiver, which is of low complexity, but it is oblivious of the interference; and 2) the ML-optimum multiuser receiver, which is of relatively high complexity, and has the benefit of being interference aware. It is shown that the single-user receiver’s performance is limited by an error floor, while the multiuser receiver has no error floor. These results confirm that both data modulation and multiple access are possible by relying only on the differences of the TA/RA channel impulse responses. Furthermore, the authors show that the error floor of the single-user receiver can be significantly reduced by increasing the number of RAs. In particular, if the number of RAs goes to infinity, the error probability goes asymptotically to zero. This behavior is today known as “massive” MIMO effect [41]. As far as the single-user receiver is concerned, the mathematical framework of [223] highlights the following trends for transmission over independent Rayleigh fading channels: 1) SS–MIMO is capable of outperforming PSK/QAM-based single-antenna communications, only if the number of TAs is higher than four, and the performance gain of SS–MIMO exponentially increases with the number of RAs. This implies that SS–MIMO should be used for rates higher than 2 bpcu per user; furthermore, 2) GSS–MIMO never outperforms SS–MIMO communications; in fact, the performance gap increases with the number of active TA elements. This result is independent of the specific choice of the spatial-constellation diagram, and it implies that the number of active TAs should in fact be minimized. As far as multiuser receivers are concerned, the mathematical framework of [223] shows that each user is inflicted a performance degradation due to the multiple-access interference, even though no error floor is present. For a generic user, this performance discrepancy increases with both the number of users and with the number of TAs, while it decreases with the number of RAs. Finally, the performance comparison with single-antenna PSK/QAM transmission shows that SS–MIMO is capable of significantly improving the performance, provided that a sufficiently high number of TA elements is available. The framework of [223] has recently also been generalized to SM–MIMO communications in [226].

In [225], Yang proposes an agile antenna-hopping spatial-division multiple-access scheme, which exploits antenna hopping with the twofold objective of providing both transmit-diversity gains and multiple-access capabilities. The associated philosophy is to transmit the same SM–MIMO symbol both over several time slots and from different TAs under the control of an appropriate antenna-hopping pattern. A class of antenna-hopping patterns obtained from fast-frequency-hopping transmission is proposed for supporting multiple-access communications and for achieving transmit diversity. The author shows analytically that the achievable diversity order of the ML-optimum demodulator increases linearly with both the number of time slots and with the number of RAs. Furthermore, the author introduces both linear and nonlinear detection schemes and investigates their achievable performance as well as complexity by simulations. The author provides precise guidelines for the design of the best demodulator as a function of the number of TAs, RAs, and of
the number of time slots in order to strike an attractive tradeoff between the performance achieved and the decoding complexity imposed.

Some papers have also investigated the attainable performance of SM–MIMO communications in the presence of multiple-access interference for transmission over dispersive fading channels [217], [219], [220]. These papers are discussed in Section IV-G.

I. Noncoherent Demodulation

In Section IV-F, it is shown that SM–MIMO communications relying on coherent demodulation is quite robust to CSIR. However, pilot symbols are required for estimating the channel impulse responses of each wireless link. Even though it is shown in [62] that the number of pilot symbols per TA does not have to be very large, nevertheless SM–MIMO faces two main challenges: 1) it relies on large-scale MIMO configurations using a large number of TAs in order to increase the throughput; as such, the total number of pilot symbols required for estimating all the channel impulse responses may be quite large; and 2) it relies on using as few RF chains as possible in order to reduce both the power consumption and the complexity imposed at the transmitter. As such, the channel impulse response of all the TA elements cannot be estimated simultaneously, which is in contrast to conventional MIMO communications, where the number of RF chains is equal to that of the TA elements. This implies that CE may require a long time, which is disadvantageous in high-mobility environments. Additionally, the results of [210] have shown that avoiding the estimation of the channel phase may result in a poor demodulation performance. Motivated by these considerations, numerous researchers have investigated the design of noncoherent demodulation schemes conceived for SM–MIMO communications [103], [154], [215], [227], [228]. Further details about the training overhead for the estimation of the CSIR and CSIT are available in Section VI-B.

Sugiura et al. [103] propose a noncoherent detection scheme for STSK–MIMO communications relying on the Cayley unitary differential encoding. The authors show that a 3-dB performance penalty is imposed compared to coherent demodulation relying on perfect CSIR, as expected. This scheme is further investigated in [228], where Xu et al. propose a demodulator that avoids the employment of the nonlinear Cayley transform, thus facilitating the employment of arbitrary complex-valued constellations. These solutions employ multiple RF branches and simultaneous symbol transmissions at the transmitter. In [154], this multi-RF constraint associated with the simultaneous transmissions is eliminated, and the authors illustrate that it is possible to activate a single TA in every symbol interval, while avoiding CSI estimation at the receiver.

In [227], a semiblind scheme is conceived for STSK–MIMO systems, where CE and data detection are performed jointly. This approach aims for using the minimum number of STSK training blocks, which is related to the number of TAs. These training blocks are first used for providing a rough initial least square channel estimate. Subsequently, single-stream ML-optimum data detection is carried out based on the initial estimate of the channel impulse responses, and the detected data are used for decision-directed least square CE. It is demonstrated that a relatively few iterations are sufficient for approaching the ML-optimum performance that can be achieved under the ideal conditions of having perfect CSIR. In [215], a similar joint channel and data detection scheme is proposed and studied under the assumption that only a single-RF chain is available at the transmitter.

J. Channel-Coded Transmission and Iterative Decoding

Practical communication systems typically employ powerful channel coding schemes, such as turbo and low-density parity-check codes, which are usually detected by using iterative or turbo decoding. In general, the performance of iteratively detected channel-encoded MIMO systems exhibits different trends in comparison to their uncoded counterparts. In this context, it is of particular importance to jointly optimize the channel encoder and the bit-to-symbol mapper for the sake of achieving area-capacity performance. Motivated by this consideration, researchers have investigated the benefits of combining SM–MIMO communications with powerful channel coding and have conceived low-complexity iterative decoding algorithms that account for the SM demodulator [55], [103], [106], [169], [229]–[234].

To elaborate a little further, in [55], Jeganathan et al. combine SSK–MIMO with BICM and propose iterative decoding exchanging extrinsic soft information with both convolutional and turbo codes. The decoder first computes the a posteriori logarithm likelihood ratios (LLRs) of the transmitted bits, which are then input to a channel decoder. The channel decoder’s LLRs representing both the information and the parity bits are then processed as extrinsic information in successive iterations of the demodulator’s LLR output. These steps are repeated until a satisfactory level of reliability and, hence, a low error probability is obtained. The authors show that substantial performance gain may be expected by using the BICM–SSK–MIMO scheme compared to single-antenna systems using BICM. Recently, Koka and Sari [231] have applied BICM to SM–MIMO communications and have shown that BICM can be an effective remedy against both channel fading and against the channel’s spatial correlation.

In [103] and [106], Sugiura et al. propose a three-stage serially concatenated transmitter for STSK–MIMO systems, where STSK is combined with multiple serially concatenated codes. At the transmitter, the input source bits are channel encoded by a half-rate RSC code and then they are interleaved by a random bit interleaver. The
interleaved bits are further encoded by an infinite-impulse response URC encoder, which are then interleaved by a second random interleaver. Finally, the interleaved bits are input to the STSK modulator. At the receiver, a three-stage iterative decoding algorithm is employed based on SISO decoders. The iterative decoder obeys the carefully considered activation of all component decoders, namely of the STSK demodulator, of the URC decoder, as well as of the RSC decoder. By using EXIT chart analysis [235], the system parameters are optimized, and it is shown that the decoder is capable of achieving near-capacity performance.

In [229], a soft-decision aided low-complexity decoder is proposed for STSK–MIMO systems. The algorithm is based on a vector-by-vector-based detector and uses a MAP symbol probability search algorithm, which exhibits a lower complexity compared to the classic block-by-block-based detector. Simulation results confirm that near-capacity performance can be obtained by combining the proposed vector-by-vector-based detector with IRCCs and URCs.

A novel reduced-complexity soft-decision-aided detector is conceived in [232] for STSK–MIMO systems, which is capable of closely approaching the performance of the optimal MAP detector. More specifically, the authors exploit a hybrid combination of a modified matched filtering concept and of a reduced-complexity exhaustive search for the sake of reducing the MAP decoding algorithm’s complexity. Furthermore, the detector is extended to support the class of generalized STSK–MIMO arrangements in [92]. Monte Carlo simulations confirm that the proposed reduced-complexity detector imposes a significantly lower complexity than the MAP detector, while inflicting only a marginal performance degradation, which is in the range of 1–2 dB. Furthermore, the authors invoke the MCMC algorithm for the sake of achieving a further complexity reduction.

Finally, the development of coded SM–MIMO communications relying on iterative decoding in the presence of multiple-access interference for transmission over dispersive fading channels is considered in [217], [219], and [220], which is discussed in Section IV-G.

V. PART III—SM–MIMO: PROMISING APPLICATION AREAS

In Section IV, the latest contributions on SM–MIMO research are summarized with emphasis on the associated physical-layer functionalities, such as encoding and decoding, along with the analysis of the information-theoretic and communication-theoretic performance in a wide variety of realistic propagation environments. The aim of this (and the next) section is to highlight that the peculiarities of SM–MIMO transmission and, in particular, the data-driven antenna-hopping encoding technique, may be applied to diverse topical research areas. More specifically, in this section, we focus our attention on those application scenarios where SM–MIMO techniques have been applied and encouraging results have been reported. Nevertheless, for the sake of projecting a realistic image, we also summarize the associated limitations and discuss a range of open research problems. On the other hand, we defer the discussion of further promising application areas to Section VI, which have hitherto received limited attention from the research community working on SM–MIMO communications.

A. Green Wireless Networks: On the Bandwidth-Efficiency Versus Energy-Efficiency Tradeoff

In Sections I and II, we have provided the motivation of the potential applications of SM–MIMO communications to green wireless networks, with the main objective of conceiving a new transmission concept that is capable of supporting high data rates, while reducing the total power consumption, which accounts for both the dynamic (RF) and static (circuit) power [30]. The EE potential of SM–MIMO communications is currently being investigated within the activities of the European project GREENET in the context of energy-efficient wireless networks design [12], [236].

The potential of SM–MIMO in green wireless communications is motivated by recent results on the fundamental limitations of conventional MIMO communications in the context of EE design. Conventional MIMO techniques are capable of reducing the power consumption of wireless communications as a benefit of their inherent multiplexing and diversity gains [28]. However, these considerations on their EE potential typically consider only their RF power consumption. On the other hand, they tend to neglect the static power consumption that is needed for the circuits, e.g., the power amplifiers. In fact, various independent results have recently unveiled that conventional MIMO communications may be less power efficient than single-antenna transmission if the total power dissipated (RF and circuit) is taken into account [31]–[34], [37]–[39]. More specifically, it is shown in these papers that the power dissipation inefficiency originates from the increased amount of power dissipated by all the independent RF chains that are needed in conventional MIMO communications. The importance of considering the total power dissipated and not only the RF transmit power in the EE evaluation framework is also acknowledged in emerging communications paradigms, such as large-scale MIMO systems. More explicitly, it is pointed out in [63] that the actual power consumption associated with supporting hundreds of active TAs will be determined by the economy of scale in manufacturing hundreds of RF chains, up/down converters, analog/digital converters, etc., as well as by the development of highly parallel and dedicated hardware for baseband signal processing.

In this context, SM–MIMO transmission has the potential of representing a win–win alternative to the circuit
power consumption inefficiency of conventional MIMO systems and to the technological challenges of implementing hundreds of small, low-power RF chains and power amplifiers, as envisaged by large-scale MIMO systems. On the one hand, SM–MIMO is capable of increasing the SE of conventional MIMO communications by incorporating numerous passive TA elements without requiring extra RF chains and without significantly impacting the static energy consumption of the power amplifiers. On the other hand, SM–MIMO is also capable of reducing the RF power consumption by capitalizing on both the multiplexing and on the diversity gains originating from the TA-hopping encoding with the aid of using only a few active RF chains. Additionally, the generalized SM–MIMO transceiver proposed in Fig. 7 may be readily integrated into the current LTE-A standard and it is backward compatible with it. In fact, the transceiver of Fig. 7 employs the same space-time-coded transmission schemes as the LTE-A standard, but it requires a large number of TA elements and a fast RF switching mechanism, which are affordable requirements in the context of millimeter-wave communications [46], [47] (see Section VI-D for further details), and with the aid of currently-available RF technology [58]–[61].

Motivated by the promising potential of SM–MIMO for EE wireless communications, a range of interesting research contributions have been published, which confirm its benefits when realistic power consumption models are considered that account for the total power dissipation [49], [237]–[240].

Chang et al. [239] design an energy-efficient transmission scheme based on the SSK–MIMO principle. The key idea is that energy-efficient communications can be established by a nonequiprobable signaling alphabet, where the low-power modulated symbols are used more frequently for transmitting a given amount of information. The design of this energy-efficient communication paradigm is formulated as a convex optimization problem, where the minimum achievable average symbol power consumption is derived under specific rate, performance, and hardware constraints. More specifically, the authors formulate a general optimization problem that minimizes the energy consumption by imposing practical constraints that simultaneously include: 1) the bandwidth efficiency; 2) the performance; and 3) the hardware complexity. More particularly, as far as the latter constraint is concerned, the proposed optimization problem includes the number of RF chains, and thus the static power consumption. It is imposed based on this power model that the larger the number of RF chains, the larger the static power consumption. The numerical results demonstrate that substantial energy gains are obtained.

In [49], [237], and [240], the authors investigate the energy-efficiency potential of SM–MIMO communications by focusing their attention on the power consumption reduction occurring from using a single-RF chain, which is in contrast to conventional MIMO communications relying on multiple active RF chains. In this study, the realistic power consumption model developed within the European project EARTH [9], [32] is considered. In this power consumption model, it is assumed that the static power consumption linearly increases with the number of active RF chains. Under the so-called SOTA-2010 power consumption model [31], [32], the EE of various types of BSs (macro, micro, pico, and femto) is investigated by comparing SM to conventional MIMO communications. The EE expressed in bits/Joule is computed as the ratio between the channel capacity and the total power supply that includes both the RF and circuit power consumption. Some of the numerical results recorded for a BS equipped with four TAs \(N_t = 4\) are depicted in Fig. 10, which is reproduced from [240] with permission. Four open-loop MIMO schemes are compared:

1) SM–MIMO using a single-RF chain;
2) STBCs transmission employing 3/4 code rate [52];
3) MISO transmission designed for achieving a transmit-diversity only;
4) spatial-multiplexing aided MIMO transmission.

The numerical results clearly highlight the EE potential of SM–MIMO compared to the other MIMO transmission schemes, especially for medium throughput. For high throughput, there is an intercept point between the curves of SM–MIMO and spatial-multiplexing MIMO, which is due to the limited number of TAs (only four) that is used for SM–MIMO. Large-scale implementations of SM–MIMO relying on a large number of passive TA elements are capable of further increasing this promising EE gain.

These recent research results clearly accentuate the inherent potential of SM–MIMO communications in the context of green wireless networks. However, a significant amount of further research work is needed in order to prove the practical energy-efficiency potential of SM–MIMO. In fact, the studies conducted so far are limited to a few conventional MIMO systems, as well as to the simplest implementation of SM–MIMO. Furthermore, the analysis is restricted to single-cell scenarios, where the other-cell interference is either neglected or implicitly treated as an additional source of noise. All these assumptions, although reasonable for preliminary analysis, constitute idealized simplifying assumptions for assessing the potential power savings in the context of green multicell networks, which are expected to be interference limited and to exploit multiuser and coordinated multipoint MIMO techniques.

B. Distributed Wireless Networks: Relaying, User Cooperation, and Network Coding

Relaying and user cooperation have recently emerged as potential candidate technologies for future wireless applications and standards [20], [21]. The main idea behind cooperative communications is that some idle wireless
Nodes (relays) may help some other active wireless nodes (sources) by receiving and retransmitting some data packets on their behalf for enhancing the reception of the destination. The fundamental design objective of efficient cooperative communications is in fact to enhance the transmission of the sources, while minimizing: 1) the extra bandwidth required; and 2) the resources of the network nodes (relays) willing to assist the sources [241]. Early transmission protocols conceived for cooperative diversity-aided communications were mainly based on the repetition coding principle relying on transmissions over orthogonal channels [21]. The main limitation of these protocols is that the diversity gain is achieved at the cost of reducing the bandwidth efficiency. Thus, in the last few years, numerous solutions have been proposed for minimizing the throughput reduction of repetition-based orthogonal relaying protocols [241]. More recently, a new family of cooperative diversity protocols has been introduced for mitigating the throughput limitations of repetition-based protocols, while still introducing an affordable implementation complexity. They are known as network-coded cooperative diversity protocols [22] conceived for striking an attractive diversity versus throughput tradeoff [242]. The main idea is that multiple packets received by the same node can be jointly encoded by superimposing them in a manner reminiscent of multicode CDMA using, for example, operations in a binary or a nonbinary Galois field. They may then be transmitted during the same time slot, while still providing the same diversity gain as conventional relying [243], [244].

As far as MIMO communications are concerned, the advantages of user cooperation are noteworthy. In fact, due to the associated size, cost, power consumption, and hardware limitations, wireless transmitters may physically not be able to accommodate multiple TAs. In this context, the benefits of MIMO communications may be realized in a distributed fashion: single-antenna-aided transmitters may share their physical antennas and create an equivalent MIMO system constituted by virtual antennas, while having the same benefits as conventional MIMO systems associated with colocated TAs [20]. Furthermore, the distributed nature of the virtual MIMO scheme relieves the system from some of the limitations of conventional MIMO communications, such as the detrimental diversity-gain erosion imposed by the spatial correlation of the tightly packed antenna elements. Owing to the potential advantages of distributed MIMO implementations, researchers have investigated the possibility of amalgamating user cooperation with SM–MIMO communications [245]–[258].

Indeed, SM–MIMO and relaying may be beneficially combined for exploiting their inherent merits, while

Fig. 10. Energy-efficiency assessment of different MIMO concepts for different types of BSs: macro, micro, pico, and femto. Reproduced from [240] with permission.
circumventing their limitations, as exemplified in the following.

- **Reduced transmit power and increased diversity.** Consider a relaying-aided MIMO system for application in the downlink of cellular networks relying on fixed RSs, where both BSs and RSs are equipped with multiple antennas [259]. In this scenario, SM–MIMO transmission may be used for replacing conventional MIMOs in order to improve the attainable performance over each wireless link with the aid of low-complexity single-RF implementations. Naturally, power gains may be expected due to the reduction of the transmission distance of relay-aided communications. Furthermore, if multiple RSs are available, an improved diversity gain can also be expected.

- **Spectral- and energy-efficient distributed uplink transmission.** In Section IV, it is shown that the gain of SM–MIMO communications benchmarked against conventional single-antenna-assisted and MIMO-assisted transmissions increases with the size of the spatial-constellation diagram. As a consequence, numerous passive TA elements have to be available, which makes SM–MIMO an attractive candidate for downlink transmission. On the other hand, even though only a single-RF chain is required for transmission, equipping a pocket-sized mobile terminal with hundreds of passive TA elements may still be infeasible in the imminent future, albeit they may be accommodated in a laptop computer. In this context, a distributed implementation of SM–MIMO schemes may be beneficial in the uplink, where hundreds of mobile terminals may share their single antennas for creating a large-scale virtual MIMO system. The information bits can be encoded onto the spatial domain indices of the mobile terminals through a data-driven mobile terminal selection mechanism, similar to SM–MIMO implementations with colocated TAs.

1The interested reader is referred to Section VI-B for discussion about the training overhead associated with the implementation of SM–MIMO communications in both the downlink and the uplink.

- **Spectral- and energy-efficient relay-aided transmission.** Conventional half-duplex relaying protocols rely on exploiting that if a user is acting as an RS for another user of the network, it must delay the transmission of its own data frames [242]. This implies that the transmission of these users is delayed and that some of their power is not used for their own transmission, but altruistically assigned for assisting other users. To circumvent this problem, advanced relaying protocols based on superposition modulation/coding and NC have been proposed in [243] and [244]. These protocols allow the RS to superimpose its own data packet with those to be forwarded on behalf of the source. This may be viewed as a win–win protocol. In fact, the same channel use is jointly employed for the RS-data transmission and for the source-data relaying. As a further benefit, no excess power is required for source-data relaying. SM–MIMO is capable of offering a similar degree of freedom for the sake of improving the bandwidth efficiency of half-duplex relaying by exploiting the additional dimension provided by the spatial-constellation diagram. For example, the data of the source may be transmitted using the signal-constellation diagram, while the data of the RS may be transmitted via the spatial-constellation diagram. Thus, both data symbols may be transmitted within the same channel use hence improving both spectral and energy efficiencies.

These potential applications of the SM–MIMO concept have been investigated by numerous researchers and encouraging results have been obtained [245]–[258].

In [246], a dual-hop MIMO-aided relaying channel is considered, where the transmitter, the RS, and the destination are equipped with multiple antennas. By assuming a demodulate-and-forward relaying protocol, the authors show that the end-to-end error probability can be improved with the aid of an SM–MIMO scheme compared to single-antenna-assisted demodulate-and-forward relaying. The performance trends associated with increasing the number of TAs is similar to those of point-to-point transmission [57]. The performance of dual-hop MIMO relaying protocols is analytically studied in [254] and [255].

In [247], Sugiuura et al. propose a distributed implementation of the STSK–MIMO principle. In particular, it is assumed that the source broadcasts a modulated signal to a set of single-antenna-aided RSs. Depending on either the success or the failure of the CRC, the RSs are activated and retransmit the received data using decode-and-forward relaying. At the relay, the STSK–MIMO principle is applied for re-encoding the received data onto a dispersion vector and onto a modulated symbol. The authors show that by taking advantage of the CRC, beneficial distributed diversity gains can be obtained because the CRC-based scheme eliminates any potential relaying-induced error propagation. The authors also propose a differential implementation, which dispenses with CE. This scheme has recently been generalized in [258]. More specifically, with the aim of mitigating the throughput loss of conventional half-duplex relaying, the authors exploit SR, where the messages ripple through a chain of RSs while avoiding the relaying-induced extra interference with the aid of orthogonal CDMA spreading codes. Explicitly, to mitigate the detrimental impact of SR-induced interference and the dispersive effects of the wireless channel, they adopt a multicarrier CDMA scheme.

In [248], Yang and Aissa propose a nonorthogonal relaying protocol that exploits the SM–MIMO principle
for the sake of throughput enhancement. The protocol is constituted by two phases. In the first (listening) phase, the source node transmits its data to both a group of RSs and the destination. In the second (cooperative) phase, the source node still transmits some data to the destination. On the other hand, the RSs that have successfully decoded the data transmitted during the first phase cooperatively retransmit it by using the SM–MIMO principle. More specifically, some received bits are mapped onto the activated RSs, while the remaining bits are transmitted using conventional modulated symbols. The RSs that are activated depending on data received from the source. The authors also evaluate the capacity of the proposed protocol and compare it to state-of-the-art nonorthogonal relaying, demonstrating that an improved capacity can be obtained.

In [252], the SSK–MIMO principle is studied for the uplink of cellular networks. It is assumed that a mobile terminal is equipped with a single TA while the BS is equipped with two RAs. It is assumed that some single-antenna-aided RSs are available in the neighborhood of the source, which form a virtual MIMO system. Since the direct link between the source and destination is of low quality, it is assumed to be unavailable. Hence, the source communicates with the destination only through the RSs. Each RS is assigned a unique identifier, which allows us to exploit the distributed SSK–MIMO principle. More specifically, the source broadcasts its data packet to the available RSs. This data packet is decoded by each RS individually and each decoded symbol is compared against their own identifier. The specific RSs that demodulate the data coinciding with their own identifier become active and transmit the associated SSK symbol to the destination. In summary, the set of RSs act as a distributed spatial-constellation diagram for the source, similar to the SSK–MIMO communications concept with colocated TAs. The destination is capable of decoding the received data due to the uniqueness of the channel impulse responses of the RS-to-destination links. Under ideal operating conditions, i.e., when each RS decodes its data without errors, the data transmitted from the source activates a single RS. However, demodulation errors encountered at the RSs may result in the activation of either none or multiple relays. The main contribution of [252] is the development of the optimal demodulator to be used at the receiver, which is capable of correctly decoding the data, and, at the same time, exploiting receive diversity. The authors confirm the performance advantage of the proposed demodulator by using Monte Carlo simulations.

The distributed encoding principle of [252] is extended in [256] with the objective of improving the achievable bandwidth efficiency of half-duplex relaying. The transmission protocol is similar to that of [252], apart from a main exception. Specifically, the active relays transmit the first data available in their buffers during the second phase. This enables each relay to transmit, during the same channel use, both the data received from the source (implicitly via the bit-to-relay mapping process during the relay activation process) and its own data (explicitly via conventional modulation). The authors analytically assess the attainable diversity gain and compare the resultant performance to that of other relaying protocols, including both an NC scheme and superposition coding/modulation. The results show that the adoption of a distributed SM–MIMO scheme is indeed capable of improving the performance.

As a further advance, a bidirectional SSK–MIMO arrangement is proposed for data transmission in [253]. At the relay, physical-layer NC is applied in order to reduce the number of channel uses needed for retransmissions. The authors show with the aid of Monte Carlo simulations that the SSK–MIMO combined with NC is capable of further improving the attainable system performance. In [249], the authors address an important practical aspect of implementing distributed/cooperative SM–MIMO systems, namely that of ensuring accurate symbol-level synchronization among the cooperative relays for high-integrity demodulation. The authors propose an iterative CFO compensation algorithm and show that an attractive performance is achieved considering practical operating conditions.

As promising as they are, the results available for characterizing the SM–MIMO in distributed wireless networks require further practical verification of their merits and limitations. For example, the analysis proposed in [247] and [258] relies on using ideal CRC at the relays. The capacity analysis of [248] assumes the absence of demodulation errors at the relays. The channel quality-aware optimal demodulators of [252] and [256] are somewhat complex to be implemented in practice, while diversity-assisted but low-complexity demodulators have not as yet been proposed in the open literature. Furthermore, the achievable diversity analysis is only applicable to simple network topologies. On the other hand, low-complexity decoders and comprehensive diversity-analysis frameworks for both conventional and NC-aided relaying protocols exist in the literature [243], [244]. Moreover, there is a paucity of literature on the impact of CSI and synchronization errors. As far as the associated synchronization requirements are concerned, a promising solution for implementing distributed MIMO networks is the application of LS spreading codes, which are capable of facilitating substantial cooperative spatial diversity gains even in the presence of large synchronization errors exceeding the symbol duration [260]. The practical rollout of distributed SM–MIMO communications requires that all these open issues be addressed.

C. Visible Light Communications: LiFi SM–MIMO

Having discussed the potential merits of SM–MIMO communications, when the information is assumed to be transmitted within the RF spectrum, i.e., by means of electromagnetic waves, let us now consider
other applications. More specifically, in this section, we show that SM–MIMO may find successful application in the visible light spectrum as well, namely in the context of VLC or OWC [261].

The increasing interest in VLC mainly originates from the significant spectrum extension that it can offer compared to the rather congested RF spectrum used by state-of-the-art wireless communications systems. In fact, in addition to being free of license fees, the spectrum that the visible light offers is about 10 000 times higher than the entire RF spectrum, which is a beneficial property in an era when the RF spectrum has become a scarce commodity. Although the concept of VLC originates from 1880 (Graham Bell, the Photophone, [262]), it was revived in 1979 [263]. However, it is only with the advent of solid-state lighting and high-brightness LEDs that the idea of using light for high-speed wireless communications has gained significant interest. However, many research challenges have to be solved and the SM–MIMO transmission principle may contribute toward this beneficial goal.

The available LEDs used for illumination produce non-coherent light. This implies that: 1) the most viable modulation technique for VLC is IM, where the desired waveform is simply imposed on the instantaneous power of an optical carrier by flickering it; and 2) the most practical down-conversion technique for VLC is DD, where a photo-detector produces a current proportional to the received instantaneous power. The DD technique is simpler to be implemented than coherent detection, since it detects only the intensity of the optical wave (frequency and phase information of the optical carrier is lost). In fact, short-range indoor optical applications usually employ IM/DD modulation, as a practical transmission technique for achieving low-complexity, low-cost optical modulation and demodulation. Although efficient and practical, the adoption of IM/DD modems imposes a fundamental limitation on the achievable rate, because light sources can only handle real-valued and unipolar signals. Strictly speaking, this implies that the well-known Shannon theorem is not directly applicable in the VLC domain. The achievable information rate of optical OFDM was quantified in [264].

In contrast to VLC schemes, the digital modulation techniques typically applied in the RF domain are based on complex-valued signals. As a consequence, the restriction of using IM/DD modems results in a 50% loss of transmission bandwidth compared to RF communications. This property also renders the employment of classical MIMO techniques used in RF communications a challenge. For example, spatial multiplexing MIMO transmissions for application to VLC require computationally complex ML-based detection and iterative interference cancellation techniques. Moreover, they are unsuitable for VLC due to the fact that the information is encoded onto the light intensity and not on the electric field, as in traditional RF communications. On the other hand, IM does not suffer from destructive fading and light intensities are superimposed constructively. Thus, diversity gains may be achieved by transmitting the same signal from multiple LEDs, using a classic technique known as RC. However, multiplexing gains cannot be obtained by employing RC.

In this context, SM–MIMOs constitute a promising technique of realizing the spatial-multiplexing gains for VLC–MIMO systems with the aid of simple IM/DD modems. For example, SSK–MIMO schemes rely on encoding the information bits onto the index of a single LED. This principle perfectly suits IM/DD modems, since the receiver only has to detect the index of the active LED. In fact, no phase information is required and it is relatively straightforward to detect the power of the signals, with the aid of a photodetector. Since numerous LEDs and photodetectors can be compactly packed at the transmitter and at the receiver, respectively, SM–MIMO may be an effective and low-complexity solution for attaining significant multiplexing gains for VLC. Motivated by this inherent potential, researchers have studied the application of SM–MIMO to VLC [265]–[280].

In [269], the SM–MIMO concept for application to VLC is introduced, under the terminology of OSM. The associated principle is the same as for RF applications: the information data are mapped onto both the index of a single LED and onto a light intensity level. The attainable performance is studied by Monte Carlo simulations for transmission over MIMO channels that are modeled with the aid of ray-tracing methods. The results show that the optical MIMO beams are highly correlated if the transmitter and receiver locations are not optimized, which potentially results in a significant power penalty. The power efficiency may be improved by increasing the number of photodetectors and/or by using channel coding techniques. Conversely, it is shown that carefully aligning the LEDs and photodetectors is capable of creating nearly uncorrelated channel paths and, hence, results in a substantial system performance enhancement. The resultant aligned scheme is shown to be both power and bandwidth efficient as compared to on/off keying modulation, pulse position modulation, and pulse amplitude modulation. The impact of transmitter-to-receiver alignment is studied in [268] as well, and it is shown that aligned OSM systems significantly enhance the attainable performance compared to unaligned ones.

A low-complexity modulation scheme for application to VLC is designed in [276], which combines SSK–MIMO and PPM. The active LED is used for transmitting a PPM signal pattern, rather than a constant optical power or a unipolar amplitude modulated power level. The simulation results provided show that the scheme benefits from both the simplicity of SSK modulation and from the energy efficiency of PPM. Compared to the conventional on/off keying modulation that transmits only a single bit/symbol, the proposed scheme is capable of increasing the data rate by a substantial factor. The achievable increase in data rate
is proportional both to the number of transmitters and to the size of the signal-constellation diagram. Furthermore, the authors construct an experimental setup for the sake of measuring the dependence of the channel on the relative position of the transmitter and the receiver. These measured channel parameters are used for evaluating the system’s error probability. The proposed scheme is compared against classical SSK modulation and RC schemes, demonstrating an improved performance.

The performance of various MIMO schemes designed for VLC, including RC, spatial multiplexing, and SM, is compared in [279]. The results show that as a benefit of its diversity gain, RC is robust to diverse transmitter-to-receiver alignments. However, since RC does not provide spatial-multiplexing gains, it requires large signal-constellation sizes for providing high spectral efficiencies. By contrast, spatial multiplexing enables high data rates at the cost of an increased SNR requirement. However, to provide these gains, spatial multiplexing also requires sufficiently low correlation among the channels of the spatial streams. As a remedy, SM is shown to be more immune to the correlation among the different VLC paths compared to spatial multiplexing, while supporting a higher throughput than RC. This is not unexpected, since either a single or a low number of LED transmitters are activated at any instant. Moreover, the authors investigate the effect of power imbalance among the LEDs. It is found that power imbalance may in fact substantially improve the performance of both spatial multiplexing and of SM, as it mitigates the detrimental effects of correlation, i.e., the coincident attenuation of the different LED transmitters’ lightpaths. This becomes particularly detrimental, when the LEDs of an array are tightly packed. Furthermore, the authors show that deactivating some of the LED transmitters reduces the channel’s correlation, which is hence capable of providing improved channel conditions for spatial multiplexing and SM. Overall, the authors show that practical VLC systems may substantially benefit from employing MIMO techniques.

VI. PART IV—SM–MIMO: THE ROAD AHEAD

In this section, we elaborate on a range of scenarios, where SM–MIMO may become promising, albeit they have not as yet been addressed by the research community. Furthermore, we summarize a suite of important, open research issues that have to be addressed for the sake of fully appraising the potential advantages of SM–MIMO communications.

A. Appraising the Fundamental Tradeoffs of Single- Versus Multi-RF MIMO Designs

As described in Sections II and III, unlike conventional spatial-multiplexing MIMO, in SM–MIMO, some of the TAs remain inactive in every channel use. Hence, they are suboptimal in terms of their SE. On the other hand, this property makes them a promising candidate for those applications where the SE is not the prime requirement to be satisfied. Indeed, often a range of other performance metrics have more precedence, thus leading to fundamental design and performance tradeoffs. In particular, when activating only a single or a few TA elements, both the circuit dissipation and the RF power consumption may be reduced. This is substantiated by recent results, showing that having many active TAs in MIMO-aided cellular networks may not be energy efficient [33]. Recent SE versus EE studies of SM–MIMOs relying on practical power consumption models have also corroborated this trend [240]. These pioneering studies further inspire the research community to draw definite conclusions about the SE versus EE tradeoff of single- versus multi-RF MIMO-aided communications. First, sufficiently rich numerical results exist only for very basic SM–MIMO schemes associated with a single-RF chain. Hence, it would be beneficial to extend them, by assessing the optimal number of active and inactive TAs in diverse operating conditions. Furthermore, more realistic scenarios should be considered. So far, only the baseline point-to-point single-cell cellular network has been studied, where the effect of other-cell interference is neglected and the characteristics of different types of BSs are not taken into account. Thus, we believe that investigating the multiuser multicell cellular scenario is a mandatory step toward assessing the ultimate potential of SM–MIMO communications as an enabler of future green cellular networks, as well as to shed light on its advantages and disadvantages against other large-scale MIMO implementations [63].

B. Large-Scale Implementations: Training Overhead for the Acquisition of CSIT and CSIR

To achieve a high throughput, SM–MIMO systems need a large number of TAs. Furthermore, to reduce the implementation complexity and to improve the EE, SM–MIMO systems are configured to operate with a limited number of RF chains compared to the available TAs. In general, the lower the number of RF chains is, the higher the EE and the lower the complexity, but the larger the number of TAs needed for a given throughput constraint. As such, SM–MIMO systems would require large-scale MIMO implementations to meet the throughput requirements of next-generation cellular networks. In the context of large-scale MIMO design, an important issue to be considered for any practical deployments is the training and feedback overhead associated with the acquisition of the CSIT/CSIR [42], [63]. Because of that, researchers are currently investigating this problem [281]–[289]. A viable solution to reduce the training overhead is the adoption of a cellular network architecture based on TDD, as discussed in [283], [284], and [286]–[288]. In fact, the channel reciprocity property may be exploited in this case and the overhead related to the channel training scales linearly.
with the number of user terminals as well as it is independent of the number of antennas at the BS. Furthermore, if adequate beamforming schemes are used at the BS, the user terminals may not require CSIR. However, most existing cellular deployments are based on FDD, where the channel reciprocity property cannot be exploited and a feedback channel is needed for downlink transmission. For these reasons, researchers are currently investigating the design of FDD-based large-scale MIMO cellular systems with a limited training and feedback overhead [281], [282], [285], and [289]. In these papers, it is shown that efficient limited-feedback strategies may be designed, whose feedback overhead is less than the number of TAs. As far as the design of SM–MIMO systems is concerned, the constraint on a limited number of RF chains and its unique encoding mechanism may further increase the training overhead compared to state-of-the-art large-scale MIMO implementations. More specifically, we state the following. 1) Due to the limited number of RF chains compared to the TAs, the channel gains of all the TAs cannot be estimated simultaneously, e.g., by using orthogonal pilot symbols. A simple but suboptimal solution to this problem is to keep activated all the available TAs during the training phase, while switching off the RF chains that are not needed for payload transmission according to the SM–MIMO encoding principle. Accordingly, the transmitter has to be equipped with a number of RF chains equal to the number of TAs. However, this solution still allows us to significantly reduce the total consumed power during data transmission. Another solution is to exploit the spatial and temporal correlation among the channel impulse responses in order to reduce the training overhead, as recently proposed in [290]. 2) Due to the SM-based encoding mechanism, the CSIR may always be needed even in the downlink of TDD-based cellular architectures. A possible but suboptimal solution may be to dispense the receiver from the CSIR by using the noncoherent demodulation schemes described in Section IV-I. In general, the impact of the training overhead for the design of SM–MIMO systems has received little attention at the time of writing, despite its importance for realizing, in practice, the potential gains of SM–MIMO transmission techniques are open loop and hence they are not directly comparable to state-of-the-art multiuser MIMO communications using precoding and beamforming. In fact, multiuser MIMO systems are radically different from open-loop transmissions, since they are based upon the concept of spatial sharing of the channel among the users. In spatial multiple access, the multiuser interference is handled by the multiple antennas, which, in addition to providing per-link diversity, also provide the necessary degrees of freedom for the spatial separation of the users. To be competitive against state-of-the-art multiuser and large-scale MIMO communications, SM–MIMO research has to rapidly move forward, by abandoning the idealized simplifying point-to-point assumption, and to realistically investigate its potential in multiuser multiecell networks.

D. Millimeter-Wave Communications: The Need for Beamforming Gains

Millimeter-wave communications is a promising technology for future cellular systems, especially for the wide availability of license-free spectrum [44], [46], [47]. Thanks to the decrease in wavelength at these frequencies, large antenna-arrays can be packed compactly at both the transmitter and the receiver. As a consequence, the millimeter-wave frequency band may be deemed a good candidate for SM–MIMO communications, since a large multiplexing gain may potentially be realized at a low complexity and at a low-power consumption, by relying on a limited number of RF chains. Enabling millimeter-wave cellular systems in practice, however, requires appropriately dealing with the channel impairments and propagation characteristics of the high-frequency bands. More specifically, the main propagation-related obstacles in realizing millimeter-wave cellular are that the free-space path loss is much larger due to the higher carrier frequency, scattering is less significant hence reducing the available diversity, and non-line-of-sight paths are weaker making blockage and coverage holes more pronounced. Further, the noise power is larger due to the use of larger bandwidth channels. As a result, for achieving an adequate SNR at the receiver, future millimeter-wave cellular networks must leverage high-gain electronically steerable directional antennas, which may be realized by beamforming or precoding data on large antenna arrays. However, some practical constraints on the complexity of baseband
signal processing and on the number of RF chains that is possible to use at these frequencies do exist, thus making the design and optimization of millimeter-wave cellular a challenge [73], [75]. In the light of these considerations, the application of SM–MIMO to the millimeter-wave frequency band needs the development of practical solutions, which can simultaneously guarantee a high multiplexing gain and a high beamforming gain, in order to offset the increased path loss and noise power. In the design and optimization spaces of SM–MIMOs, this introduces another important constraint: the number of RF chains cannot be arbitrarily low in order to provide the necessary high beamforming gain. The number of RF chains has to be appropriately chosen by taking into account the required beamforming gain, the power consumption/dissipation of the power amplifiers, the implementation complexity of baseband signal processing, and the need for low-complexity RF-based precoding schemes, as well as the required throughput. Furthermore, these design and optimization problems have to be studied by considering realistic propagation channel models for the millimeter-wave frequency band, which may differ significantly from conventional channel models [46], [73]. At the time of writing, the application of SM–MIMO to millimeter-wave frequencies has not as yet been investigated. Due to the large number of TAs that can be accommodated at these frequency bands, however, important practical contributions may be made in this unexplored research area.

E. Small-Cell Heterogeneous Cellular Networks: Toward Interference Engineering

Cellular networks are undergoing a major shift in their deployment and optimization [291], [292]. New infrastructure elements, such as femto/pico BSs, fixed/mobile relays, cognitive radios, unlicensed millimeter-wave radios, and distributed antennas are being deployed, thus making future cellular systems more heterogeneous [293]. As a consequence, new cellular deployments are characterized by a more unplanned, irregular and random location of many infrastructure elements, whose positions may vary widely. As a result, the interference patterns are becoming more complex making the evaluation and design of different communication technologies and protocols more challenging [294]. Therefore, accurately modeling the interference produced by randomly deployed network elements, understanding the impact on the system’s performance and developing efficient techniques for mitigating it are three fundamental research facets. Furthermore, this new heterogeneous cellular topology poses fundamental questions about the best way of exploiting the access points and antenna elements. Raising dilemmas, such as using a few BSs in combination with large-scale MIMOs, using many small cells illuminated by low-power BSs having a few TA elements each, or whether to combine large-scale MIMOs and small-cell networks, have to be addressed. In this context, SM–MIMO introduces another dimension to be investigated: the tradeoff between the number of active and inactive TA elements in each access point, in order to strike the desired SE versus EE tradeoff. Furthermore, the employment of low-power BSs has the beneficial effect of reducing the transmission distance between mobile terminals and access points, which results in performance improvements. This reduction of the transmission distance may be favorable to SM–MIMO communications, since the static power consumption may have a more pronounced impact compared to the RF power consumption. The success of the SM–MIMO transmission principle depends on its efficiency, flexibility, and performance that it will be able to offer in this emerging heterogeneous cellular scenario. Some preliminary results about the assessment of SM–MIMO communications in the presence of heterogeneous interference have recently appeared in [295] and [296]. The authors capitalize on research advances in network interference modeling using stochastic geometry. However, these results are not directly applicable to cellular networks, since the cell association mechanism is not considered and only the impact of random interference is investigated. Owing to its unique working principle, SM–MIMO requires accurate interference models for ensuring that the interference is carefully managed for the sake of approaching the optimum performance. Thus, interference-aware transmitter and receiver designs are necessary.

F. RF Energy Harvesting: Taking Advantage of the Idle Antennas

In Sections II and III, we have extensively commented on the necessity of future cellular networks to become more energy efficient. We have also shown that SM–MIMO constitutes a promising enabler of reducing the total power consumption of cellular networks, by deactivating some TAs. However, other solutions may be conceived for exploiting the available TAs and for improving the EE of cellular networks. An option that is currently gaining momentum in green networking is RF energy harvesting, which can provide additional energy supply for wireless devices [297]. The main idea is that RF signals may be used for transporting both information and energy simultaneously. In fact, the ambient RF radiation can be captured by the available RA and may be converted into a direct-current voltage simply using appropriate circuits. In a cellular network, RF energy harvesting may be realized using new hybrid cellular deployments, where some stations often referred to as power beacons are overlaid onto the existing cellular network with the objective of charging wireless devices with the aid of microwave radiation links relying on microwave power transfer [298], [299]. In this context, SM–MIMO may be a suitable transmission technology for taking advantage of the RF energy harvesting principle and of the related shift in the topology of cellular networks for enabling joint information and power transfer. On the one hand, SM–MIMO
inherently reduces the interference compared to conventional MIMOs, since many TAs are inactive. This reduction of the interference results in a better SIR, which, in turn, reduces the transmit power of the other radios in the network, hence increasing the overall EE. Furthermore, deactivating some TAs reduces the circuit power consumption quite considerably. On the other hand, the EE may also be improved by dispensing with switching off the TAs that are idle, instead switching them to receive mode. This way, the TAs that are not used for data transmission may be used for RF energy harvesting by capturing ambient RF radiation and converting it into power that can be used for subsequent transmissions. Since the number of available antennas may be high in SM–MIMO systems, all the inactive TAs may be switched to RAs for the sake of capturing a nonnegligible amount of RF power. Of course, appropriate transceivers have to be implemented for enabling the simultaneous transmission of information and the reception of power in unison with the symbol time switching mechanism of SM–MIMOs [300]. Although hitherto completely unexplored, the amalgamation of SM–MIMO and RF energy harvesting may be a promising research area for sustainable low-power networking.

G. Leveraging the Antenna Modulation Principle to a Larger Extent

Throughout this paper, we have advocated a new way of transmitting data by exploiting the unique and location-specific channel impulse responses associated with different TA elements. More specifically, SM–MIMO maps some information bits onto the indices of the TAs. This modulation technique is efficient, since data demodulation is facilitated by exploiting the differences in the channel impulse responses of the TAs. Motivated by the SM–MIMO principle, some fundamental questions naturally arise: Could other characteristics of the antenna array be used for data modulation? May they be combined for the sake of realizing high multiplexing gains at a low complexity, while relying on a limited number of RF chains for striking a flexible SE versus EE tradeoff? We believe that the answers to these questions are affirmative and that the SM–MIMO principle is only a single specific example of exploiting some unique characteristics of the antenna arrays for transmitting information. Let us deepen this concept a little further.

In Section IV-B, many transmitter designs closely related to the SM–MIMO concept have been described, which do not encode the information bits onto the indices of the TAs. For example, the dispersion matrix aided approach of [103] maps the information onto the index of legitimate dispersion matrices. The larger the number of these matrices is, the higher the multiplexing gain, albeit this is achieved at the cost of an increased detection complexity. In [139], an antenna-pattern modulation scheme is introduced, where the bits are mapped onto the TA indices and onto a set of legitimate radiation patterns. In [40], the information bits are mapped onto orthogonal bases defined in the beam-space domain. Other options may include encoding information bits onto different antenna polarizations, onto beamforming vectors or onto the diverse combination of all the modulation techniques just mentioned. We believe that the promising achievable performance and low-complexity implementation of the original SM–MIMO concept suggests that every unique characteristic (or their combinations) of the antenna array may be exploited for data modulation, provided that the antennas may be appropriately designed and these uniquely distinguishable antenna features may be extracted at the receiver side. This approach to data modulation relies upon the design of new antenna arrays and it goes beyond the common idea of simply feeding different TA elements with diverse bit streams for the sake of increasing the achievable throughput. On the other hand, multiple data streams may be used for choosing appropriate “features” of the TA array and for transmitting a considerable amount of information bits without creating ICI.

H. Open Physical-Layer Research Issues

In Section VI-C, we have emphasized the urgent need for shifting the focus of SM–MIMO research from point-to-point to multiuser, multicell scenarios for assessing its advantages in comparison to the current trends in MIMO research. However, fundamental key aspects at the physical layer have not as yet been addressed. In this context, a nonexhaustive list of open research issues may be formulated as follows.

- The design of single-RF SM–MIMO systems simultaneously providing both transmit diversity and ML-optimum single-stream decoding complexity remains an open issue at the time of writing. The proposal in [50] still requires multiple RF chains and the design of adequate shaping filters for low-complexity detection. The proposal in [203] does provide a single-RF solution, but its low-complexity decoding as well as its comparison to state-of-the-art MIMO is still unexplored. The proposal in [103] relies on the numerical design of appropriate dispersion matrices, but this task is nontrivial for a large number of TAs. This problem is partially addressed in [149], where, however, many active TA elements may be needed for achieving good performance. Furthermore, the amalgamation of multicarrier LTE-style transceivers with the single-RF SM–MIMO architecture is still a fairly open research issue.

- By relying on a large number of TA elements compared to the number of RF chains, SM–MIMOs are unable to estimate all the channel impulse responses of the antenna array at the same time. This implies that the training overhead required for CE may increase with the ratio of the number of TAs and the number of available RF chains.
Although SM–MIMOs have been shown to be robust to CE errors and some noncoherent demodulators can dispense with CE at the receiver, demodulation schemes relying on coherent detection require efficient CE algorithms for large-scale implementations.

- It is apparent from our state-of-the-art survey that using a fixed number of TAs and MIMO transmission modes regardless of the fading channel is a suboptimal solution in terms of both spectral- and energy-efficiency performance metrics. On the other hand, link-adaptive SM–MIMO schemes that adapt the number of active/inactive TAs and the number of information streams to be transmitted constitute a more promising option. Indeed, link-adaptive SM–MIMO communications have been proposed in [128]. However, the techniques advocated require a nonnegligible feedback overhead and impose a high computational complexity. Furthermore, their achievable performance has not been analytically characterized, hence it has to be assessed through extensive simulations.

- The application of SM–MIMOs to distributed and relay-aided networks implies that adequate demodulation schemes have to be available at the destination for taking advantage of the inherent diversity gain offered by multirelay transmission. The receivers available at the time of writing are computational demanding and rely on the assumption of idealized perfect CSI. In this context, new receiver architectures that are less complex but are robust to the knowledge of the CSI have to be developed. Furthermore, accurate symbol-level synchronization may be needed to fully benefit from the spatial-constellation diagram. Quasi-synchronous distributed designs may be investigated to this end [260].

- The heterogeneous and ad hoc nature of future cellular networks requires transceivers that are no longer oblivious of the interference, but exploit different grades of coordination among the randomly distributed network elements, as well as the statistical characteristics of the aggregate interference for conceiving improved transmission and reception algorithms. Interference-aware SM–MIMO encoding and decoding schemes have not as yet been designed. Furthermore, the distribution of the aggregate other-cell interference of SM–MIMO cellular communications is still unknown. Advanced mathematical tools borrowed from the area of applied probability, such as stochastic geometry, may be used to this end [292].

I. Implementation Challenges of SM–MIMO Design

In Section VII, we will summarize a recent activity related to the experimental testbed-based assessment of SM–MIMO communications. The experiments carried out to date have confirmed that the information bits mapped onto the TA indices can be decoded with the expected performance in both outdoor and indoor environments. However, a range of further important implementation aspects related to the design of SM–MIMOss have to be solved for realizing the full potential of single-RF-based large-scale MIMO implementations. Some important implementation issues are as follows.

- The single-RF implementation requires a fast RF switch. The switching has to be performed at the symbol rate and low insertion losses have to be guaranteed. Furthermore, the switch needs to be energy efficient for its application to power-efficient cellular networks.

- New large-scale antenna-array architectures have to be designed for enhancing the uniqueness of the channel signatures. Furthermore, reconfigurable antenna-array designs are needed for generalized modulation schemes that are capable of exploiting the features of the available TAs. Both mutual coupling and electromagnetic compatibility properties of these TAs have to be investigated.

- Switching at the symbol rate requires shaping filter designs that are capable of striking a beneficial tradeoff between the practical bandwidth and the time duration in order to avoid leakage losses and, at the same time, for complying with regulatory spectral emission masks.

- Combining single-RF MIMO designs and multicarrier/filterbank modulation may necessitate sophisticated RF and baseband signal processing techniques.

VII. PART V—SM–MIMO: FROM THEORY TO PRACTICE

In this section, we describe the world’s first experimental assessment of SM–MIMO transmission with the aid of a testbed platform deployed in an indoor propagation environment. The results illustrated in this section are in part reproduced from [94] with permission.

A. Transmission Chain

The transmission chain used in the testbed setup is shown in Fig. 11, which is constituted by hardware and software parts. The hardware consists of a National Instruments (NI)-PXIe-1075 chassis at the transmitter (PXIe-Tx) and another NI-PXIe-1075 chassis at the receiver (PXIe-Rx), each equipped with the relevant NI modules [301]. The software carries out the digital signal processing at the transmitter (DSP-Tx) and at the receiver (DSP-Rx). At the transmitter, the binary data are processed by the DSP-Tx block before being transmitted through the channel by the PXIe-Tx. At the receiver, the PXIe-Rx records the RF signal...
and outputs it to the DSP-Rx, where the original data stream is recovered.

B. Hardware Testbed

Both the NI-PXIe-1075 chassis of the transmitter and of the receiver are equipped with a 1.8-GHz Intel-i7 processor relying on 4-GB RAM. The system has two TAs and two RAs. Each TA and RA contains two quarter-wave dipoles, and a half-wave dipole placed in the middle. All three dipoles are vertically polarized. Additionally, each antenna has a peak gain of 7 dBi in the azimuth plane, with an omnidirectional radiation pattern. The signals are transmitted at a carrier frequency of 2.3 GHz.

The PXIe-Tx transmitter hardware is composed of three interconnected hardware modules:

1) the NI-PXIe-5450 I/Q signal generator, which is fed with the transmit vector extracted from the binary file generated in MATLAB using the DSP-Tx encoding algorithm; the DSP-Tx module is described in Section VII-C;
2) the NI-PXIe-5652 RF signal generator, which is connected to the NI-PXIe-5611 frequency converter, which takes its input signal from the output of the NI-PXIe-5450 I/Q signal generator;
3) the NI-PXIe-5611 frequency converter, which outputs the analog waveform corresponding to the binary data at a carrier frequency of 2.3 GHz.

Likewise, the PXIe-Rx receiver hardware is composed of three interconnected hardware modules:

1) the NI-PXIe-5601 RF downconverter, which is used for detecting the analog RF signal from the RAs;
2) the NI-PXIe-5652 onboard reference clock, which is used for synchronization;
3) the NI-PXIe-5622 IF digitizer, which applies its own bandpass filter and produces the received binary files that are processed in MATLAB by the DSP-Rx decoding algorithm; the DSP-Rx module is described in Section VII-C.

C. Software Testbed

MATLAB is used for the digital signal processing required both at the transmitter and the receiver, DSP-Tx and DSP-Rx, respectively. DSP-Tx processes the incoming information data and generates a file that can be transmitted by PXIe-Tx. DSP-Rx processes the data received by PXIe-Rx and recovers the original binary data stream. The processing algorithms at DSP-Tx and DSP-Rx are shown in Fig. 12.

The following operations are implemented at the DSP-Tx.

1) The binary data are first split into information segments of appropriate size.
2) The information in each segment is then modulated using SM-MIMO.
3) A pilot signal is incorporated for the sake of CE along with a frequency offset estimation section.
4) Then, zero-padding is performed, which permits up-sampling of the data, while maintaining the same signal power. The up-sampling ratio is set to four and the up-sampled data are passed through a root raised cosine finite impulse response filter with 40 taps and a rolloff factor of 0.75. A large rolloff factor and a long tap delay are necessary for ensuring that the power is focused to a short time window, i.e., for ensuring that only a single antenna is activated in every channel use.
5) The resultant vector is multiplied by a factor termed as the “tuning signal power” for the sake of adjusting the desired transmit power required for the information sequence.
6) The frames are created such that the frame length multiplied by the sampling rate is less than the coherence time of the channel, which is typically about 7 ms for a stationary indoor environment. This ensures that all channel estimates generated at the receiver are valid for the frame duration.

The following operations are implemented at DSP-Rx.

1) The raw data vectors received from the NI-PXIe-5622 digitizer are combined, in order to form the received matrix for data demodulation.
2) The detector first finds the beginning of the transmitted sequence by using the synchronization sequence (based on an autocorrelation algorithm).
3) The SNR for each vector is calculated using the “SNR section.”
4) Each raw vector is decomposed into its underlying frames.
5) Each frame is down-sampled and passed through a root raised cosine filter which completes the matched filtering.
6) Frequency offset estimation, timing recovery, and correction of each frame follow, which rely on using state-of-the-art algorithms.
7) The pilot signal is then used for CE.
8) The remaining data, along with the estimated channels, are finally used for recovering the estimated binary sequence (ML-optimum demodulation).

D. Experimental Evaluation

The physical layout of the experimental setup is shown in Fig. 13, and the relative antenna spacing is provided in Fig. 14. More particularly, the two TAs and two RAs are identical and are placed directly across each other. As such, the channel between the transmitter and the receiver has a strong LoS component. Therefore, the transmit-to-receive channel exhibits a Rician distribution with a high Rician factor due to the short distance between the transmitter and the receiver. In order to confirm the Rician distribution and to estimate the Rician factor, some channel measurements are collected and studied with the help of the chi-squared goodness-of-fit test. These measurements confirm that the channel is Rician distributed with a Rician factor in the range of 31–38 dB.

To assess the performance of the testbed, some measurements are performed and the error probability is computed experimentally. Specifically, a stream of $10^4$ information bits is transmitted. These data are split into 50 frames with 200 bits each. To implement coherent detection, the channel is estimated both at the beginning and at the end of every frame, resulting in 100 CEs per transmission. The experiment is repeated 1000 times for every SNR point and the error probability is computed. The results are shown in Fig. 15. A good agreement between the experiments and both the Monte Carlo simulations (“Sim”) and the analytical results (“Ana”) can be observed. Explicitly, the analytical results are obtained by considering a Rician distribution with a Rician factor equal to 33 dB and by taking into account some of the associated hardware imperfections, such as the power imbalance of the RF chains. The gap observed at low SNR can be attributed to a number of practical factors, including the imperfect frequency offset estimation, timing recovery errors, synchronization problems, and inaccurate CE.

In summary, the preliminary results illustrated in this section for a simple $(2 \times 2)$-element MIMO testbed, including full transmit and reception chains and signal processing algorithms, confirm that the theoretical gain predicted by the analysis and simulations are substantiated by measurements performed in a controlled laboratory environment.
VIII. CONCLUSION

The need for power-efficient MIMO-aided cellular networks requires a paradigm shift in the wireless system design. This trend is irreversible and will have a profound impact on both the theory and practice of future heterogeneous cellular networks, which will no longer be purely optimized for approaching the attainable capacity, but will explicitly include the energy efficiency during the design and optimization of the entire protocol stack. In this paper, we have critically appraised SM, which constitutes a promising transmission concept in the context of MIMO communications, and have described both a business case and the technical foundations for making it a suitable air-interface candidate for power-efficient, yet low-complexity MIMO cellular networks.

We have conjectured that the SM concept can be further leveraged, by exploiting the beneficial features of large-scale antenna arrays for low-complexity transceiver designs and for energy-efficient front-end concepts at the transmitter, while relying on a limited number of RF chains. Preliminary experimental results substantiating the benefits of SM have also been illustrated. It is our hope that

Table 1 Main Takeaway Messages and Design Guidelines for SM-MIMO Systems, Which Can Be Inferred From the State-of-the-Art Survey of Sections IV and V

<table>
<thead>
<tr>
<th>Summary and Design Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rate:</strong> The achievable rate of SM-MIMO transmission depends on the number of RF chains $(N_{RF})$. The higher $N_{RF}$, the higher the achievable rate, but the higher the transmitter complexity and power dissipation. If $N_{RF} &gt; 1$,ICI-free transmission is still possible, provided that a single data stream is encoded onto the signal-constellation diagram and the $N_{RF}$-tuple of TA-indices constitute the spatial-constellation diagram. The rate can be further increased by multiplexing $N_{RF}$ data streams and encoding them onto the signal-constellation diagram. The price to pay is a higher receiver complexity imposed by the ICI-mitigation [306].</td>
</tr>
<tr>
<td><strong>Capacity:</strong> The mutual information of SM-MIMO systems configured with a single-RF chain ($N_{RF} = 1$) and Gaussian input signals depends on the number of TA-elements ($N_{TA}$) [89], and it is almost the same as the capacity of the so-called “spatial cycling using a single transmitter at a time” MIMO scheme [54, p. 317]. For transmission over identically distributed fading channels, the ergodic capacity is almost the same as that of single-antenna aided transmission. On the other hand, the outage capacity increases with the number of TA-elements $N_{TA}$.</td>
</tr>
<tr>
<td><strong>Error Performance:</strong> For transmission over independent and identically distributed Rayleigh fading channels, SM-MIMO transmission is capable of providing a better ABEP than single-antenna assisted transmission if $N_{T} &gt; 4$ and the number of RA-elements is $N_{R} &gt; 1$ [62], [57]. The higher the number of RA-elements, the higher the gain compared to single-antenna transmission. In general, SM-MIMO transmission takes better advantage of receiver diversity than single-antenna transmission [57]. For transmission over independent and identically distributed multiple-access Rayleigh fading channels, the multi-user interference asymptotically vanishes if $N_{U} \rightarrow \infty$ [223], which is referred to in parlance as the “massive MIMO effect”. SM-MIMOs enhanced with space-time coding are capable of outperforming single-antenna transmission and conventional space-time-coded MIMO schemes even for $N_{T} = 1$, despite relying on low-complexity single-stream detection at the receiver [50].</td>
</tr>
<tr>
<td><strong>Channel Fading:</strong> The specific distribution of the channel fading significantly affects the achievable diversity gain and the associated BER of SM-MIMO transmissions [57]. For transmission over benign Nakagami and Rician fading channels, the less dramatic the fading, the higher the Rician and Nakagami fading factor and hence more bits should be encoded onto the signal-constellation diagram [57].</td>
</tr>
<tr>
<td><strong>Channel State Information:</strong> SM-MIMO transmission is robust to the imperfect knowledge of CSIR. If adequate channel estimators are used, SM-MIMOs are more robust to channel estimation errors than conventional MIMO transmission schemes. If low-complexity designs are sought, differentially encoded and non-coherently detected schemes constitute an attractive design option at the cost of a modest performance erosion.</td>
</tr>
<tr>
<td><strong>Transmit-Preprocessing:</strong> The performance of SM-MIMOs can be enhanced if CSIT is available through a feedback channel. Both signal- and spatial-constellation diagrams can be optimized and a wide range of pre-processing algorithms can be designed by trading-off performance for feedback overhead and complexity [99], [128], [140].</td>
</tr>
<tr>
<td><strong>Demodulation:</strong> A wide range of demodulation schemes are available for striking a flexible trade-off between the achievable performance and detection complexity imposed, ranging from the matched-filter that provides the worst performance at the lowest complexity [53] and the ML-optimum decoder that provides the best performance at the highest complexity [56]. Flexible demodulation schemes include near-optimum sphere decoding [155], low-complexity schemes based on compressed sensing [165], and enhanced matched-filter aided demodulators [154]. The best demodulator to use can be chosen in the light of the complexity constraints of the receiver.</td>
</tr>
<tr>
<td><strong>Energy-Efficiency:</strong> SM-MIMO transmission constitutes a promising power-efficient MIMO scheme for the following two reasons. 1) It is capable of reducing the total power dissipation of the power amplifiers by taking advantage of a single-RF implementation. 2) It is capable of reducing the RF transmit-power by taking advantage of the multiplexing gain offered by the spatial-constellation diagram. By using realistic power consumption models, SM-MIMO transmissions were shown to provide a better energy-efficiency than conventional MIMO schemes for medium/high rates [240]. The number of RF chains can be optimized for striking a flexible trade-off between the attainable throughput and the total power consumption.</td>
</tr>
</tbody>
</table>
these promising results will inspire more research on SM in the years to come.

We close this paper by directing interested readers to a collection of companion tutorial slides on SM research [295] and to some YouTube videos, where the theory and practice of SM are described [302]–[304]. Furthermore, a two-hour video presentation that introduces this paper is available on YouTube [305]. The slides of the video presentation are available for download as well [306]. Finally, we summarize in Table 1 the main takeaway messages and design guidelines for SM–MIMO systems, which can be inferred from the state-of-the-art survey of Sections IV and V.

Acknowledgment

M. Di Renzo and H. Haas would like to acknowledge colleagues and students participating in the European project GREENET (Grant 264759), where part of the ideas and results presented in this paper have been developed. Special thanks go to A. Younis, N. Serafimovski, T. Fath, A. Stavridis, X. Wu, S. Narayanan, K. Ntontin, and D. De Leonardis for their long-lasting effort on SM–MIMO research. M. Di Renzo would like to thank C. Verikoukis and the Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain, for giving him the opportunity of recording the companion video of [305] in CTTC’s facilities in Barcelona, Spain.

REFERENCES


[67] YB power wireless system antenna number, 3GPP TSG-RAN WG2 #67, contribution R2-094677 from Huawei.


Di Renzo et al.: Spatial Modulation for Generalized MIMO: Challenges, Opportunities, and Implementation


[308] SMPaper_Slides.pdf

Marco Di Renzo (Member, IEEE) was born in L’Aquila, Italy, in 1978. He received the Laurea (cum laude) and Ph.D. degrees in electrical and information engineering from the Department of Electrical and Information Engineering, University of L’Aquila, L’Aquila, Italy, in April 2003 and January 2007, respectively, and the Habilitation à Diriger des Recherches (HDR) from the University of Paris-Sud XI, Paris, France, in October 2013.

From August 2002 to January 2008, he was with the Center of Excellence for Research DEWS, University of L’Aquila. From February 2008 to April 2009, he was a Research Associate with the Telecommunications Technological Center of Catalonia (CTTC), Barcelona, Spain. From May 2009 to December 2009, he was an EPSRC Research Fellow with the Institute for Digital Communications (IDCOM), University of Edinburgh, Edinburgh, U.K. Since January 2010, he has been a Tenured Researcher (“Chargé de Recherche Titulaire”) with the French National Center for Scientific Research (CNRS), as well as a faculty member of the Laboratory of Signals and Systems (L2S), a joint research laboratory of the CNRS, the École Supérieure d’Électricité (SUPELEC), and the University of Paris-Sud XI, GIF-sur-Yvette, France. His main research interests are in the area of wireless communications theory. He is a Principal Investigator of four European-funded research projects (Marie Curie ITN-GREENET, Marie Curie IAPP-WSN4QOL, Marie Curie ITN-CROSSFIRE, and Marie Curie IAPP-SmartNRG).

Dr. Di Renzo is the recipient of a special mention for the outstanding five-year (1997-2003) academic career, University of L’Aquila, Italy; the THALES Communications fellowship for doctoral studies (2003-2006), University of L’Aquila, Italy; the Best Spin-Off Company Award (2004), Abruzzo Region, Italy; the Torres Quevedo award for research on ultrawideband systems and cooperative localization for wireless networks (2008-2009), Ministry of Science and Innovation, Spain; the “Dérogation pour l’Encadrement de Thèse” (2010), University of Paris-Sud XI, France; the 2012 IEEE CAMAD Best Paper Award from the IEEE Communications Society; the 2012 Exemplary Reviewer Award from the IEEE WIRELESS COMMUNICATIONS LETTERS of the IEEE Communications Society; the 2013 IEEE VTC-Fall Best Student Paper Award from the IEEE Vehicular Technology Society; and the 2013 NoE-NEWCOM# Best Paper Award. He currently serves as an Editor of the IEEE COMMUNICATIONS LETTERS.

Harald Haas (Member, IEEE) received the Dipl.-Ing. degree from the University of Applied Sciences, Nuremberg, Germany, in 1994, and the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K., in 2001.

He holds the Chair of Mobile Communications in the Institute for Digital Communications (IDCOM), University of Edinburgh, Edinburgh, U.K., and he currently is the CTO of a university spin-off company pureVLC Ltd. His main research interests are in interference coordination in wireless networks, spatial modulation, and optical wireless communication. He holds 23 patents. He has published more than 60 journal papers, including a Science article and more than 160 peer-reviewed conference papers. Nine of his papers were invited papers. He has coauthored a book entitled Next Generation Mobile Access Technologies: Implementing TDD (Cambridge, U.K.: Cambridge Univ. Press, 2012). Since 2007, he has been a Regular High Level Visiting Scientist supported by the Chinese 111-program at Beijing University of Posts and Telecommunications (BUPT), Beijing, China.

Prof. Haas was an invited speaker at the 2011 TED Global Conference, and his work on optical wireless communication was listed among the “50 best inventions in 2011” in the Time Magazine. He recently has been awarded a prestigious Fellowship of the Engineering and Physical Sciences Research Council (EPSRC) in the United Kingdom.

Ali Ghrayeb (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Arizona, Tucson, AZ, USA, in 2000.

He is currently a Professor with the Department of Electrical and Computer Engineering, Texas A&M University at Qatar, Doha, Qatar (on leave from Concordia University, Montreal, QC, Canada). He is the coauthor of the book Coding for MIMO Communication Systems (New York, NY, USA: Wiley, 2008). His research interests include wireless and mobile communications, error-correcting coding, MIMO systems, wireless cooperative networks, and cognitive radio systems.

Dr. Ghrayeb is a corecipient of the 2010 IEEE Global Communications Conference (Globecom) Best Paper Award. He has instructed/coconstructed technical tutorials at several major IEEE conferences. He served as the TPC Cochair of the Communications Theory Symposium of IEEE Globecom 2011. He is serving as the TPC Cochair of the 2016 IEEE Wireless Communications and Networking Conference (WCNC). He serves as an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON COMMUNICATIONS. He served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, Elsevier Physical Communications, and the Wiley Wireless Communications and Mobile Computing Journal.

Shinya Sugihara (Senior Member, IEEE) received the B.S. and M.S. degrees in aeronautics and astronautics from Kyoto University, Kyoto, Japan, in 2002 and 2004, respectively, and the Ph.D. degree in electronics and electrical engineering from the University of Southampton, Southampton, U.K., in 2010.

From 2004 to 2012, he was a Research Scientist with Toyota Central Research and Development Laboratories, Inc., Tokyo, Japan. Since January 2013, he has been an Associate Professor with the Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Tokyo, Japan, where he heads the Wireless Communications Research Group. His research has covered a range of areas in wireless communications, networking, signal processing, and antenna design. He has authored or coauthored over 50 refereed research publications, including 30 IEEE journal and magazine papers.

Dr. Sugihara was the recipient of numerous awards, including the 28th Telecom System Technology Award from the Telecommunications Advancement Foundation in 2013, the 6th IEEE Communications Society Asia-Pacific Outstanding Young Researcher Award in 2011, the 13th Ericsson Young Scientist Award in 2011, and the 2008 IEEE Antennas and Propagation Society Japan Chapter Young Engineer Award. He was also certified as an Exemplary Reviewer of IEEE COMMUNICATIONS LETTERS in 2013.
Lajos Hanzo (Fellow, IEEE) received the first-class degree in electronics in 1976 and the Ph.D. degree from the Technical University of Budapest, Hungary, in 1983. He was awarded the Doctor of Sciences (D.Sc.) degree by the University of Southampton, Southampton, U.K., in 2004. He is a Fellow of the Royal Academy of Engineering. During his career in telecommunications, he held various research and academic posts in Hungary, Germany, and the United Kingdom. Since 1986, he has been with the Department of Electronics and Computer Science, University of Southampton, where he holds the Chair in Telecommunications. He coauthored 20 Wiley and IEEE Press books, totaling in excess of 10 000 pages on mobile radio communications and published 1380 research papers and book chapters on IEEE Xplore. He is also active in organizing and chairing IEEE conferences, such as the 2006 and 2009 Wireless Communications and Networking Conference (WCNC), the 2010 Spring and Fall and 2011 Spring Vehicular Technology Conference (VTC), the 2013 IEEE International Conference on Communications (ICC), etc. He also presented keynote and overview lectures and has been awarded a number of distinctions. Currently, he heads a 100-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by the industry, the Engineering and Physical Sciences Research Council (EPSRC) U.K., the European FP 7 Programme, and the Mobile Virtual Centre of Excellence (VCE), U.K. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also an IEEE Distinguished Lecturer and a Governor of the Vehicular Technology Society, a Fellow of the IET/IEEE. He is an editorial board member of the PROCEEDINGS OF THE IEEE. During 2008–2012, he was the Editor-in-Chief of the IEEE Press.