Enabling Technologies for Spectrum and Energy Efficient NOMA-MmWave-MaMIMO Systems

Yue Wang, Zhi Tian, and Xiuzhen Cheng

Abstract

With the proliferation of versatile devices and data-consuming services, the quest for spectrum efficiency has led to the merging of three disruptive technologies: millimeter-wave (mmWave) communications, massive MIMO (maMIMO), and non-orthogonal multiple access (NOMA). Emerging wireless networks move toward ultra-dense deployment of massive devices with diverse service demands, which call for efficient spectrum sharing even on mmWave bands. This article studies some key techniques that account for the unique angular selectivity of mmWave maMIMO channels and thus enable re-engineering the spectrum sharing paradigm of NOMA. An overview is provided on research challenges and opportunities related to spectrum and energy efficiency of spectrum-shared NOMA-mmWave systems with maMIMO, with focus on high-performance and low-complexity channel sensing, optimal sensing resource allocation, security and privacy provisioning, and learning-aided real-time system optimization.

Introduction

The gamut of wireless services for various applications over valuable radio frequency (RF) spectrum has deeply penetrated our modern society. The quest for spectrum efficiency has led to the merging of three disruptive technologies: millimeter-wave (mmWave) communications, massive multiple-input multiple-output (maMIMO), and non-orthogonal multiple access (NOMA). Recent advances in integrated circuit technologies make mmWave systems cost-effective for practical deployment, tapping into a large amount of underutilized bandwidth in the over-30 GHz bands to support multi-gigabit-per-second wireless access. Given the abundant bandwidth, current efforts on mmWave focus on making use of maMIMO for reliable and energy-efficient transmission in the unique propagation environments, without much concern for spectrum efficiency.

Future wireless networks move toward ultra-dense deployments of massive devices with diverse service types, which will inevitably add spectrum efficiency as a major consideration. Indeed, mmWave communications are envisioned to support low-power high-density applications including machine-to-machine and the Internet of Things, where a large number of devices within a small area need to be connected seamlessly. As such, needs for spectrum sharing arise even on mmWave bands in order to connect massive devices. NOMA, via power-domain or code-domain multiplexing, allows devices to share the same spectrum and time resources. The combination leads to NOMA-mmWave-maMIMO communications with high spectrum efficiency, as documented by capacity analyses [1]. NOMA-mmWave-maMIMO is well motivated to fulfill the requirements of high data rate, low latency, and massive connectivity.

For NOMA, the benefits in spectrum efficiency come at potential costs in energy efficiency. For instance, random beamforming in mmWave NOMA requires increased power margin [2]. To improve energy efficiency, great efforts have to be spent on user grouping, power allocation, and beamforming, suggesting various multi-beam and beamspace schemes in the physical layer (PHY). Noticeably, the benefits of nearly all these PHY solutions in NOMA hinge on accurate channel knowledge, in the form of either the channel state information (CSI) or some key channel parameters such as angles of arrival/departure (AoA/AoD) and path gains. However, high-performance channel sensing entails large energy consumption and long sensing time, especially for densely deployed and spectrum-sharing mmWave NOMA, and such challenges are aggravated for maMIMO due to the large antenna size. Either a long sensing time or a non-negligible channel estimation error adversely affects the effective data rates, and hence may offset the spectrum efficiency offered by NOMA. Evidently, efficient channel sensing plays a key role in unleashing the envisioned spectrum and energy efficiency of NOMA-mmWave-maMIMO systems.

This article aims to provide an overview on efficient sensing techniques for large-size, multi-dimensional, and directional channels, focusing on those with high sample efficiency, estimation accuracy, computational efficiency, and applicability to practical NOMA-mmWave-maMIMO transceiver architectures. The most recent advances on efficient super-resolution channel sensing are investigated, followed by discussions on practical implementations and theoretical guarantees. Further, the tradeoff between channel sensing and data transmission is illuminated to suggest optimal resource allocation. Finally, related research issues and future directions are highlighted.

Yue Wang and Zhi Tian are with George Mason University; Xiuzhen Cheng is with George Washington University.

Digital Object Identifier: 10.1109/MWC.001.2000055

1536-1284/20/$25.00 © 2020 IEEE

The intertwined functional modules of NOMA-mmWave-maMIMO hinge on accurate knowledge of user-specific channels, where users within a group may experience similar propagation channels that cannot be distinguished unless accompanied by super-resolution sensing techniques. The paradigm of maMIMO-enabled NOMA-mmWave faces a severe bottleneck in terms of large energy consumption and long sensing time, and hence urgently requires novel efficient channel sensing techniques.

Bottleneck of NOMA-MmWave-MaMIMO

NOMA-mmWave-maMIMO systems aim to fulfill the increasing demands for spectrum sharing, massive connectivity, ultra-dense deployment, heterogeneous data traffic, high bandwidth efficiency, and ultra-reliable and ultra-low latency services. Therein, each base station (BS) is equipped with large-scale antenna arrays to provide various services to its physically adjacent users that are closely located in a small area. Via NOMA, the BS employs user-specific beamformers and power levels to send superimposed messages to a group of users in the downlink, which involves several functional modules to control the intra-cell interference: power allocation, precoding and beamforming, user scheduling and grouping, and user ordering for successive interference cancellation (SIC) [1, 2].

As shown in Fig. 1, all these intertwined functional modules of NOMA-mmWave-maMIMO hinge on accurate knowledge of user-specific channels to retain the self-term data and remove the interference terms caused by adjacent users. Due to NOMA, users within a group may experience similar propagation channels that cannot be distinguished unless accompanied by super-resolution sensing techniques. The communication viewpoint of channel estimation is to directly estimate the CSI from the received training signal. However, the huge antenna number in maMIMO poses formidable challenges to channel estimation. Not only do the signal acquisition and hardware costs increase drastically, but also the processing complexity and sample costs of conventional channel estimation techniques become prohibitively high. Evidently, the paradigm of maMIMO-enabled NOMA-mmWave faces a severe bottleneck in terms of large energy consumption and long sensing time, and hence urgently requires novel efficient channel sensing techniques.

Meanwhile, signal propagation over mmWave bands is highly directional in the angular domain [3]. This directional channel characteristic suggests an array processing framework, but has to cope with low sample efficiency and may cause the pilot contamination issue, especially in the presence of large antenna arrays. Advances in virtual channel modeling and sparse channel sensing reap high sample efficiency [4, 5], but suffer from limited angular resolution and degraded accuracy due to the on-grid assumption of compressed sensing (CS). Recent progress on super-resolution channel sensing via gridless CS offers high sensing accuracy at high sample efficiency, but relies on idealized array geometry with uniform antenna spacing [6–8]. In view of these opportunities and obstacles, this article focuses on the critical issues and key techniques in enhancing the sensing efficiency, facilitating the trade-off in sensing resource allocation, and promoting real-time system optimization.

In an all-connected future world, an unprecedented amount of private and sensitive data is transmitted over wireless networks, underpinning the importance of security and privacy. Being a spectrum-shared paradigm, NOMA-mmWave-maMIMO systems can be more vulnerable to eavesdroppers and privacy breach than conventional non-NOMA wireless networks. Hence, it is essential to impose requirements on transmission security guarantee and data privacy protection in NOMA-mmWave-maMIMO systems.

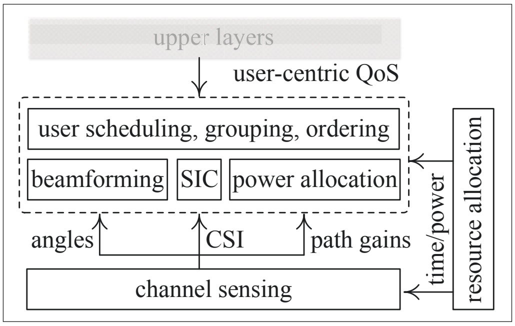

FIGURE 1. Intertwined functional modules of NOMA-mmWave-maMIMO where channel sensing serves as the prerequisite and foundation.

Sparse Channel and Sensing Model

To facilitate accurate user-specific beamforming, channel sensing for NOMA-mmWave-maMIMO focuses on the challenges incurred by large-scale antenna arrays at the BS. Consider a basic setup of fixed wireless scenarios where the BS is equipped with $N \gg 1$ antennas, in the form of either 1D uniform linear array (ULA) or 2D uniform planar array (UPA). Each user has an M-element ULA (MIMO), which can reduce to a single antenna with $M = 1$ (single-input multiple-output, SIMO) for low-cost devices. The BS performs uplink channel estimation, and the well-recognized channel or angle reciprocity is invoked to allow beamforming in downlink transmission [3].

MmWave propagation experiences limited scattering with sparse multipath, which induces unique angular directionality of mmWave maMIMO channels [3]. Specifically, the channel can be modeled as $\mathbf{H} = \Sigma_{l=1}^{L} \alpha_l \mathbf{a}_N(\theta_l)\mathbf{a}_M^H(\psi_l)$, which consists of a small number of L channel paths, each parameterized by a path gain α_l and path angles (θ_l, ψ_l) indicating AoA and AoD. The manifold vectors $\mathbf{a}_N(\theta_l)$ and $\mathbf{a}_M(\psi_l)$ reflect the array geometry. This 2D model subsumes either the MIMO case or SIMO with a UPA of size $(N \times M)$ at the BS. When the BS employs ULA and the user has a single antenna, the channel reduces to the simplest 1D case, $\mathbf{h} = \Sigma_{l=1}^{L} \alpha_l \mathbf{a}_N(\theta_l)$.

When such mmWave channel structures are ignored, conventional channel estimators turn out to be ineffective for maMIMO in terms of both energy and sample efficiency. A remedy is to take an angle-based viewpoint to reshape and enhance the transceiver design of NOMA-mmWave-maMIMO systems. Given the sparse parametric channel modeling, the task of channel sensing boils down to estimating the channel path parameters $\{\theta_l, \psi_l, \alpha_l\}_l$. The perplexing channel estimation task is accordingly broken down into intertwined subtasks that can be solved with high sample and computing efficiency.

Angle Estimation via Subspace Methods

The traditional array processing viewpoint suggests acquiring users' angular directions using super-resolution angle estimators including multiple signal classification (MUSIC), estimation of signal parameters via rational invariance techniques (ESPRIT), maximum likelihood (ML), and so on. These techniques work effectively provided that an adequate number of snapshots are available to well approximate the signal covariance via sample averaging. Unfortunate-

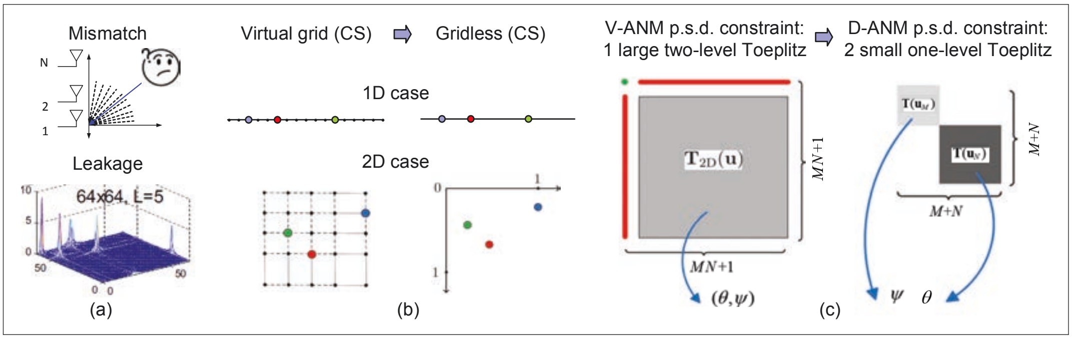

FIGURE 2. Motivation and basic ideas of gridless CS: a) grid mismatch and power leakage issues in traditional on-grid CS due to the finite grid size; b) extending on-grid CS to gridless CS in both 1D and 2D cases; c) decoupling high-dimensional ANM problem into lower dimensions, as reflected in the structure of the p.s.d. constraints of V-ANM vs. D-ANM.

ly, such a requirement causes low energy efficiency for maMIMO with large antenna arrays.

To solve this problem, smoothing techniques can be incorporated into subspace methods, but at the cost of reduced aperture size and hence degraded resolution, which is counterproductive. Besides, subspace methods usually rely on certain prior knowledge of the underlying signal, such as the number of paths, which is unknown or hard to anticipate accurately in practice.

Channel Sensing via CS

Capitalizing on the multipath sparsity of mmWave channels, recent developments seek to reduce the sampling burden and training overhead by sensing a sparse channel from compressive observations over a small number of snapshots, through virtual MIMO channel representation and CS techniques [4, 5]. The basic idea is to take the discrete Fourier transform matrix as the sparsifying dictionary and apply the CS principle for sparse channel recovery. It amounts to assuming that the AoAs/AoDs reside exactly on some fixed virtual grids in the angular domain [5], which simplifies angle estimation and yields high sample efficiency.

However, this on-grid CS channel sensing approach offers limited angular resolution due to the finite grid size. When signals arrive off-grid, the grid mismatch issue arises, which causes power leakage, as shown in Fig. 2a, and results in an error floor even in the high signal-to-noise ratio (SNR) region. It is inadequate to meet the high-accuracy requirements for channel or angle estimation, and hence cannot fulfill the needs for highly directional beamforming and effective interference mitigation in NOMA-mmWave-maMIMO spectrum sharing.

Channel angular directionality can also be calibrated by CS-based beam training techniques. However, beam training takes time to cover a wide angular region and is user-specific with synchronized beaming. Hence, the training overhead scales linearly with the number of users, which is too high for NOMA-based dense networks. These practical issues limit applicability of beam training for NOMA-mmWave spectrum sharing.

Super-Resolution Channel Sensing via Gridless CS

To overcome the grid mismatch issue of conventional grid-based CS, gridless CS techniques are proposed as sparse signal processing using a sparsifying dictionary of infinite size [6–8], extending the concept of on-grid CS to allow continuous signal locations, as illustrated in Fig. 2b.

1D Case

Let us start from a simple 1D channel **h**, which is composed of a few components from a known atom set $\mathcal{A} = \{\mathbf{a}_N(\theta), \forall\theta \in [-\pi/2, \pi/2]\}$ of infinite size, but the composition is unknown. In fact, for a given vector **h**, its decomposition over $\mathcal{A}$ is not unique. It is asserted that under certain conditions on source separation, the sparsest atomic decomposition of **h** over $\mathcal{A}$ yielding the atomic norm $\|\mathbf{h}\|_{\mathcal{A}}$ is indeed the true decomposition of **h** [6]. The sought atomic norm minimization (ANM) offers super-resolution, since θ is off-grid in $\mathcal{A}$.

ANM involves infinite programming, but can be reformulated into a tractable form if the atom set obeys some desired geometric structures, such as the Vandermonde structure in $\mathbf{a}_N(\theta)$ of ULA. In that case, ANM can be reformulated via semidefinite programming (SDP), based on the fact that any low-rank, positive semidefinite (p.s.d.) Toeplitz matrix allows unique Vandermonde decomposition [6]. Given received **y** and pilot symbol *s*, an ANM-based channel estimator arises [7]:

$$\min_{\mathbf{h},\mathbf{u},\mu} \|\mathbf{y}-\mathbf{h}s\|_2^2 + \frac{\lambda}{2}(\text{trace}\,(\mathbf{T}(\mathbf{u}))+\mu)$$

$$\text{s.t.} \begin{bmatrix} \mu & \mathbf{h}^H \\ \mathbf{h} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq 0,$$

where **u** is the first row of a Toeplitz matrix **T** to be optimized. At the optimal $\mathbf{u}^\blacklozenge$, the Vandermonde decomposition of $\mathbf{T}(\mathbf{u}^\blacklozenge)$ yields the desired angles that form **h**. It estimates both the channel **h** and angles $\{\theta_l\}$ at super-resolution, from a single measurement vector **y**. Such salient features are desired to circumvent the power leakage and pilot contamination issues in maMIMO. On the other hand, the computational complexity of SDP-based ANM is $O((N + K)^{3.5D})$ [8], where *K* is the number of snapshots dictating the sensing time, and *D* is the dimensionality, provided that the Vandermonde structure is present in all dimensions. The benefits of gridless CS appear to be limited to small *K* and *D* = 1.

2D Case

For a 2D channel **H** with *D* = 2, a straightforward extension from the 1D case is to vectorize **H** into a long vector of length *NM* corresponding to a vec-

To overcome the grid mismatch issue of conventional grid-based CS, gridless CS techniques via ANM are proposed as sparse signal processing using a sparsifying dictionary of infinite size, extending the concept of on-grid CS to allow continuous signal locations. To overcome the computational challenge of vectorized-ANM in 2D cases, most recently advances have been made in developing efficient 2D gridless CS that retains all the benefits of ANM while remarkably reducing the computational complexity through efficient reformulation of 2D ANM.

While ANM-based gridless CS offers evident performance benefits to maMIMO channel sensing, it is of practical significance to develop low-complexity implementations in order to further reduce the energy consumption in computation. To this end, a fast algorithm named iterative Vandermonde decomposition and shrinkage-thresholding (IVDST) is developed by using the accelerated proximal gradient technique, which significantly reduces the overall computational complexity compared to SDP-based ANM.

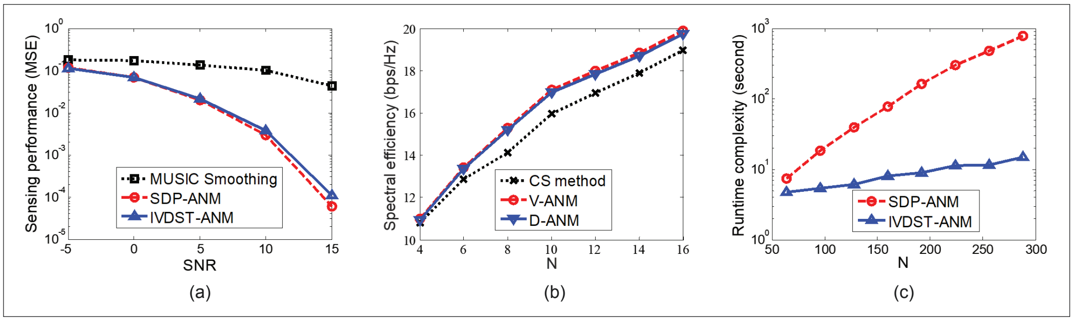

FIGURE 3. Gridless CS for $N \times N$ channel sensing: a) estimate accuracy; b) spectral efficiency; c) runtime complexity.

tor-form atom set $\mathcal{A}_v = \{\mathbf{a}_M^*(\psi) \otimes \mathbf{a}_N(\theta), \forall(\theta, \psi) \in [-\pi/2, \pi/2]\}$, and then formulate a similar SDP via a two-level Toeplitz structured matrix $\mathbf{T}_{2D}(\mathbf{u})$ [7]. However, the computational order of this vectorized-ANM (V-ANM) reaches $O(N^{3.5}M^{3.5})$ mostly because of the increased size of the p.s.d. constraint in the SDP, which becomes computationally infeasible for large N and/or M and in turn causes impractical energy consumption and hardware costs.

To overcome this challenge, most recently, advances have been made in developing efficient 2D gridless CS that retains all the benefits of ANM while remarkably reducing the computational complexity through efficient reformulation of 2D ANM [8]. The key novelty is to introduce a new matrix-form atom set $\mathcal{A}_d = \{\mathbf{a}_N(\theta)\mathbf{a}_M^H(\psi), \forall(\theta, \psi) \in [-\pi/2, \pi/2]\}$, which naturally decouples the high-dimensional ANM problem into lower dimensions and at the same time retains joint 2D optimality in the formulation [8]:

$$\min_{\mathbf{H}, \mathbf{u}_M, \mathbf{u}_N} \|\mathbf{y} - \mathbf{H}s\|_2^2 + \frac{\lambda}{2}(\text{trace }(\mathbf{T}(\mathbf{u}_M)) + \text{trace }(\mathbf{T}(\mathbf{u}_N)))$$

$$\text{s.t.} \begin{bmatrix} \mathbf{T}(\mathbf{u}_M) & \mathbf{H}^H \\ \mathbf{H} & \mathbf{T}(\mathbf{u}_N) \end{bmatrix} \geq 0.$$

This decoupled-ANM (D-ANM) is expressed by two decoupled one-level Toeplitz matrices $\mathbf{T}(\mathbf{u}_M)$ and $\mathbf{T}(\mathbf{u}_N)$ in the SDP, as illustrated in Fig. 2c. Because of the structural decoupling in D-ANM, the computation along the two dimensions is maximally decoupled without loss of joint optimality, which reduces the complexity order to $O((N + M)^{3.5})$. This novel solution comes with rigorous analysis on its identifiability conditions and theoretical sample complexity [8].

Practice and Theory

In NOMA-mmWave-maMIMO, channel sensing has to cope with several practical challenges imposed by the need for fast algorithm implementation and by various hardware constraints. It is also crucial to analyze the fundamental limits of channel sensing under practical constraints in terms of sample efficiency and performance bounds.

Fast Algorithms

While ANM-based gridless CS offers evident performance benefits to maMIMO channel sensing, it is of practical significance to develop low-complexity implementations in order to further reduce the energy consumption in computation. To this end, a fast algorithm named iterative Vandermonde decomposition and shrinkage-thresholding (IVDST) is developed, which reduces the overall computational complexity to be on the order of $O((N + M)^2)$ by using the accelerated proximal gradient technique in lieu of SDP [9]. In each iteration of IVDST, a Vandermonde decomposition is applied to utilize the Toeplitz structure inherent in the array geometry. Meanwhile, to approximate the proximal operator, the low-rank property of the Toeplitz-structured matrix is enforced via a simple shrinkage-thresholding operation. The IVDST algorithm offers an explicit way to bridge the ANM principle with classic super-resolution angle estimation algorithms [9]. Figure 3 corroborates that ANM through fast algorithm implementation achieves high channel sensing accuracy and high spectral efficiency at low runtime complexity.

Hybrid Precoding Constraints

Practical NOMA-mmWave-maMIMO systems typically adopt a hybrid analog-digital transceiver architecture, in order to balance between array gain and energy consumption. As shown in Fig. 4, it keeps a large antenna size N at the analog front-end, but employs only a few digital circuits in the form of N_{RF} ($\ll N$) RF chains. During the channel estimation stage with unknown angles, one can use a random codebook to form the precoding matrix $\mathbf{W} = \mathbf{W}_{\text{RF}}\mathbf{W}_{BB}$, where entries of the analog precoder $\mathbf{W}_{RF}$ are typically restricted to have constant modulus, for instance, via phase shifters. Such hardware limitations directly impact the ANM formulation and the theoretical sample efficiency, both of which have been addressed in [10]. Essentially, the least-squares term in the objective needs to be adjusted in accordance to the specific hybrid structure, and the p.s.d. constraint needs to be carefully reformulated in order to reflect the geometric structure of the hybrid beamformer while keeping the computing cost low.

Arbitrary or Imperfect Arrays

The success of gridless CS via SDP-based ANM is largely due to its ingenuous use of the special array geometric structure, namely Vandermonde manifold, which arises naturally from ULA and UPA. However, as shown in Fig. 4, practical NOMA-mmWave-maMIMO systems may encounter the following non-ideal array geometries that render SDP inapplicable: ULA or UPA with perturbation due to array mismatch, arbitrary array geometry, and antenna selection (from ULA, UPA, or an arbitrary array) for reduced complexity during channel estimation. A remedy in coping with these practical issues of imperfect array geometries has been developed in [10], which stems

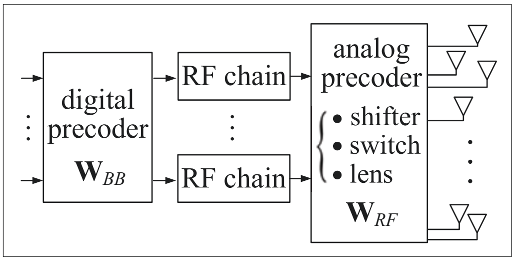

FIGURE 4. Hybrid precoder hardware with arbitrary arrays.

from the array manifold separation techniques in array processing. It approximates a length-N array manifold vector of any geometry $\mathbf{p}$ and any frequency f by means of $\mathbf{a}_N(\theta; \mathbf{p}, f) = \mathbf{B}(\mathbf{p}; f)\mathbf{v}_{\tilde{N}}(\theta)$, where $\mathbf{B}(\mathbf{p}; f)$ is derived from an $\tilde{N}$-term Bessel or Fourier approximation given $\mathbf{p}$ and f, and $\mathbf{v}_{\tilde{N}}(\theta)$ is a virtual ULA-like array manifold of length-$\tilde{N}$ and thus has the Vandermonde structure. Because the transformed new atom set $\tilde{\mathcal{A}} = \{\mathbf{v}_{\tilde{N}}(\theta), \forall\theta\}$ always possesses the desired Vandermonde structure, gridless CS becomes feasible based on $\tilde{\mathcal{A}}$ [10].

Fundamental Limits

Parallel to the development of channel and angle estimators, the associated fundamental limits need to be delineated in order to assess the sample efficiency and estimation accuracy. Analysis of this nature has been studied for D-ANM in terms of achievable compression ratio [8]. Extending such results to practical NOMA-mmWave-maMIMO settings is quite meaningful. For instance, under the hybrid maMIMO architecture, the precoding matrix $\mathbf{W}_{RF}$ of a constrained structure would affect both the feasibility condition and the sample efficiency of ANM, which informs a trade-off in hardware complexity and sensing time. When the number of RF chains is reduced to lower the power cost, the mutual incoherence property of $\mathbf{W}$ may become less strong, which may lead to longer sensing time [10]. Meanwhile, in the manifold separation approach for arbitrary arrays, the length $\tilde{N}$ of virtual array reflects the trade-off between approximation errors, spatial resolution, and computational complexity [10]. Further, the analysis on Cramér-Rao lower bounds has been performed to understand the impact of key parameters on the sensing errors of gridless CS [8].

Having reviewed the design principles, practical considerations, and fundamental limits of super-resolution sensing based on gridless CS, this section ends with a comprehensive comparison of various categories of channel sensing techniques, as summarized in Table 1.

Sensing Resource Allocation

The overall efficiency of NOMA-mmWave-maMIMO transceivers relies critically on judicious allocation of transmission resources, a portion of which should be used to ensure successful channel sensing. Channel estimation accuracy depends on not only the precoding mechanisms and channel estimators, but also the sensing time and transmission power. The transmission resources used for channel sensing have to be balanced with that for conveying information messages in order to strike desirable trade-offs between sensing accuracy and information rate. To this end, this section presents a couple of meaningful directions.

First, optimizing the sensing resources boils down to selecting the power and number of pilot symbols through a resource allocation optimization problem. Let us consider block fading channels whose channel coefficients vary from block to block but stay invariant within each transmission burst of time length N_T. Each burst includes N_s pilot symbols for sensing and $N_i = N_T - N_s$ for conveying information messages. Meanwhile, a fixed total transmission energy per burst $E_T = E_s + E_i$ is allocated to pilot and information symbols, respectively. Apparently, when N_s and E_s increase, both the sensing accuracy and the resulting per-symbol detection accuracy improve, but N_i and E_i have to be reduced accordingly, which may offset the benefits. To maximize the effective information rate, it is viable to use the average capacity under CSI errors as the performance metric for resource allocation optimization [11].

Further, the channel sensing task itself can be decoupled into two steps, and sensing resources can be allocated accordingly to improve overall efficiency. Specifically, among the channel path parameters $(\theta_l, \psi_l, \alpha_l)_l$, angle parameters (θ_l, ψ_l) vary slowly and stay invariant over multiple transmission bursts, whereas the path gains α_l fade relatively fast and need to be updated for every burst. Meanwhile, accurate angle estimation and tracking take a long sensing time and heavy computing power, but can be done in a blind mode from information-bearing signals as well [4, 10]. As such, a two-step approach to channel sensing is prudent. The resource allocation problem involves determining the number of transmission bursts to be spent on the first step of angle estimation in the blind mode, while the training block within each burst can be made much smaller, only for the second step of estimating path gains given the estimated angles [10]. It is quite illuminating to compare the outcomes of the one-step vs. two-step approaches to channel sensing. Such results are useful in striking desirable performance-rate trade-offs in order to achieve maximum system capacity under total transmission resource constraints [7].

Open Issues and Opportunities

This section discusses a few open research issues in NOMA-mmWave-maMIMO with outlooks on novel technical approaches and perspectives.

High-Dimensional Wideband Channel Sensing

Channel sensing for NOMA-mmWave-maMIMO often becomes a high-dimensional parameter estimation problem. Factors adding to the channel dimensionality include the expansion to UPA and possibly cubic arrays at the BS, the deployment of ULA and even UPA at the user side, as well as possible frequency selectivity and Doppler frequency. Such problems, subsuming wideband angle estimation, are formidably challenging for SDP-based ANM. This is because not only does the complexity of ANM grow exponentially in the problem dimension D, but also the required geometric structure can be missing in some dimensions. To solve these obstacles, a key idea is to utilize the array manifold separation principle, which can provide dimensionality separation capability as well. Indeed, the frequency parameters in higher-order Bessel approximation can include both frequency bins in wideband processing, and Doppler frequencies. This idea has not been studied in the literature, and

Practical NOMA-mmWave-maMIMO systems typically adopt a hybrid analog-digital transceiver architecture in order to balance between array gain and energy consumption. Such hardware limitations directly impact the ANM formulation and the theoretical sample efficiency. Practical antenna systems may encounter the non-ideal array geometries that render SDP inapplicable. A remedy is developed based on the array manifold separation techniques for arbitrary arrays.

Aspect	Category		
	Traditional subspace	On-grid CS	Gridless CS
Estimator applied	MUSIC, ESPRIT, ML	L1-norm minimization	Atomic norm minimization
Channel properties used	Statistical structure	Sparsity	Sparsity + structured array geometry
Signal acquisition cost	High	Low	Low
Number of snapshots	Multiple	Single or multiple	Single or multiple
Sensing time	Long	Short	Short
Estimation resolution	Super-resolution	Limited resolution	Super-resolution
Computational complexity	Low	High or low via fast algorithms	High or low via fast algorithms
Hybrid precoding	Inapplicable	Applicable	Applicable
Array perturbation	Applicable or manifold separation required	Applicable	Manifold separation required

TABLE 1. A comparison among different sensing techniques.

requires deep understanding on manifold-based modeling and processing.

The complexity issue of ANM associated with problem dimensionality is further compounded by the presence of multiple measurement vectors. While the D-ANM [8], by virtue of its decoupling strategy, effectively reduces the computational complexity to be comparable to that of a 1D ANM solution at no loss of optimality, it only applies to the case of a single measurement vector.

There is still a lack of computationally efficient gridless high-dimensional channel sensing methods when multiple measurements are present. To fill such a gap, it is crucial to exploit the structured features of channel statistics for efficient high-dimensional parameter estimation [12], which leads to a new regime where statistical signal processing meets structure-based optimization.

Security and Privacy Provisioning

Research on NOMA-mmWave-maMIMO has just started to seek understanding of the system-level design and functional module development, but leaves security and privacy largely unattended.

The transmission security can be measured by secrecy capacity, defined as the rate difference between the legitimate transmitter-receiver channel and the transmitter-eavesdropper channel. There is barely any work on the secrecy of NOMA for mmWave maMIMO systems, where the channel directionality has crucial implications. Cooperative jamming, as an important way to provide secrecy, relies on CSI for optimization and demands certain robustness to CSI errors. A main approach is to generate artificial noise signal that can be eliminated at the legitimate receiver but cannot be eliminated at the eavesdroppers so as to degrade the transmitter-eavesdropper channel. Such noise signals can be generated by some friendly jammers other than the information-bearing transmitter, giving rise to the cooperative jamming approach that can work jointly with artificial noise [13]. Through NOMA, multi-access transmission can send information messages and serve the dual role of cooperative jammers at the same time. The dual role of cooperative jammers, along with the new dimension of spatial-domain multiple access for NOMA in mmWave maMIMO networks, have not been explicitly treated in the literature, and need cross-layer design with joint consideration of user fairness and upper-layer quality of service (QoS).

To enable cooperation and improve the network level spectrum and energy efficiency, the NOMA-mmWave-maMIMO systems depend on network awareness among legitimate BSs and users. As such, sensitive messages need to be exchanged among the network entities. To protect the privacy of these messages, it may prove effective to employ differentially private message exchange schemes. The idea is to add proper noise terms to the public signals by tailoring the differential privacy principle to NOMA-mmWave-maMIMO design with consideration of the following two key factors. First, not all exposed messages result in the same level of security vulnerability. For energy efficiency, it is important to identify and protect those valuable messages in the network. Second, the added noise may degrade the network performance, resembling the channel uncertainty effect. The stronger the privacy is, the more impactful it can be on the network optimization outcomes. A recent work on differentially private alternating direction method of multipliers for distributed network optimization sheds light on this important trade-off issue [14]. The trade-off analysis and resulting robust design will help to determine the operating regimes in which secure cooperation truly results in performance gains.

Learning-Aided Real-Time System Optimization

Since NOMA-mmWave-maMIMO systems confront several optimization problems for parameter/feature extraction and resource allocation, they typically require considerable computing time in order to converge to the optimal solutions. Thus, they face considerable challenges in real-time computing, which is worsened by the complexity and diversity of future networks to which NOMA-mmWave-maMIMO responds.

Recently, deep learning (DL) has attracted great attention for enhancing the performance and functionalities of wireless communications in the big data era [15]. DL provides important capabilities in implementing real-time system optimization by developing and testing a learning-aided framework for enhancing the optimization efficiency. Along this direction, a deep neural network (DNN) can be adopted to mimic an optimizer of interest in an optimization-guided DL framework, as depicted in Fig. 5. It is well appreciated that DNN offers a highly competitive solution to approximate an unknown nonlinear functional mapping between the input and output, given a sufficient set of labeled input-output data for training. For a DNN of moderate size, once it is well trained, it can process the input data quickly to yield the desired output, making it suitable for real-time processing.

The key question is how to obtain the training dataset. Domain knowledge in NOMA-mmWave-maMIMO can be used to generate the training dataset for a DNN. Further, a network simulator can be used to generate the network conditions and input data for training. In addition, to improve the learning accuracy with high sample efficiency, a hybrid-DNN that combines model-driven optimization and data-driven learning can be designed by redefining certain layers of the DNN as signal processing layers; for instance, a noise-suppressing matched filter with tunable filter parameters can be embedded at the first layer of a DNN to autonomously enhance signal and data cleansing. Last but not least, when DL is introduced to NOMA-mmWave-maMIMO systems, there is a need to perform joint resource allocation to both learning and communication modules in order to enhance the overall spectrum and energy efficiency. This intricate trade-off is unique to learning-aided system optimization for wireless communications and has not been well investigated in the literature to the best of our knowledge.

Summary

NOMA-mmWave-maMIMO epitomizes an important multi-technology aggregation for spectrum- and energy-efficient wireless communications. This article reviews several key enabling techniques, and discusses the challenges and opportunities in NOMA-mmWave-maMIMO. Inspired by the unique propagation characteristics, sparse parametric modeling for channel sensing is provided, followed by highlights on various techniques for accurate and efficient channel sensing as well as directions for optimal sensing resource allocation. Future topics and potential research trends are also envisioned, related to high-dimensional wideband channel sensing, security and privacy provisioning, and learning-aided real-time system optimization for NOMA-mmWave-maMIMO.

Acknowledgment

This work was partly supported by the U.S. National Science Foundation (NSF) grants #1527396, #1547364, #1704397, and #1851197.

References

[1] D. Zhang *et al.*, "Capacity Analysis of NOMA with Mmwave Massive MIMO Systems," *IEEE JSAC*, vol. 35, no. 7, July 2017, pp. 1606–18.

[2] Z. Ding, P. Fan, and H. V. Poor, "Random Beamforming in Millimeterwave NOMA Networks," *IEEE Access*, vol. 5, 2017, pp. 7667–81.

[3] T. S. Rappaport *et al.*, *Millimeter Wave Wireless Communications*, Pearson Education, 2014.

[4] Y. Wang *et al.*, "Efficient Channel Statistics Estimation for Millimeter-Wave MIMO Systems," *Proc. IEEE Int'l. Conf. Acoustics, Speech and Signal Processing*, Shanghai, China, Mar. 2016, pp. 3411–15.

[5] Y. Wang *et al.*, "A Fast Channel Estimation Approach for Millimeter-Wave Massive MIMO Systems," *Proc. IEEE Global Conf. Signal and Info. Processing*, Washington, DC, Dec. 2016, pp. 1413–17.

[6] G. Tang *et al.*, "Compressed Sensing Off the Grid," *IEEE Trans. Info. Theory*, vol. 59, no. 11, Nov. 2013, pp. 7465–90.

[7] Y. Wang, P. Xu, and Z. Tian, "Efficient Channel Estimation for Massive MIMO Systems via Truncated Two-Dimensional Atomic Norm Minimization," *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–6.

[8] Z. Zhang, Y. Wang, and Z. Tian, "Efficient Two-Dimensional Line Spectrum Estimation Based on Decoupled Atomic Norm Minimization," *Signal Processing*, vol. 163, Oct. 2019, pp. 95–106.

[9] Y. Wang and Z. Tian, "IVDST: A Fast Algorithm for Atomic Norm Minimization in Line Spectral Estimation," *IEEE Signal Processing Letters*, vol. 25, no. 11, Nov. 2018, pp. 1715–19.

[10] Y. Wang *et al.*, "Super-Resolution Channel Estimation for Arbitrary Arrays in Hybrid Millimeter-Wave Massive MIMO Systems," *IEEE J. Selected Topics in Signal Processing*, vol. 13, no. 5, Sept. 2019, pp. 947–60.

[11] Z. Tian and G. B. Giannakis, "A GLRT Approach to Data-Aided Timing Acquisition in UWB Radios-Part II: Training Sequence Design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, Nov. 2005, pp. 2994–3004.

[12] Y. Wang *et al.*, "Efficient Superresolution Two-Dimensional Harmonic Retrieval via Enhanced Low-Rank Structured Covariance Reconstruction," *Proc. IEEE Int'l. Conf. Acoustics, Speech and Signal Processing*, Barcelona, Spain, May 2020, pp. 5720–24.

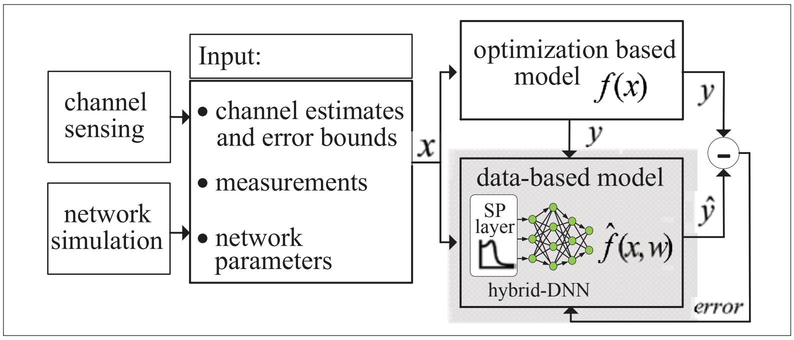

FIGURE 5. Optimization-guided deep learning.

[13] Y. Huo *et al.*, "Secure Communications in Tiered 5G Wireless Networks with Cooperative Jamming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, June 2019, pp. 3265–80.

[14] X. Cao *et al.*, "Differentially Private ADMM for Regularized Consensus Optimization," *IEEE Trans. Automatic Control*, to appear 2020.

[15] Y. Wang and Z. Tian, "Big Data in 5G," *Encyclopedia of Wireless Networks*, Springer, Mar. 2018; http://dx.doi.org/10.1007/978-3-319-32903-1_58-1

Biographies

Yue Wang [M'11] (ywang56@gmu.edu) received his Ph.D. degree in communication and information system from Beijing University of Posts and Telecommunications, China, in 2011. He is currently a postdoctoral researcher with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, Virginia. Prior to that, he was a senior engineer with Huawei Technologies Co., Ltd., Beijing, China. From 2009 to 2011, he was a visiting Ph.D. student with the Department of Electrical and Computer Engineering, Michigan Technological University, Houghton. His general interests lie in the areas of signal processing, wireless communications, artificial intelligence, and their applications in cyber physical systems. His current research focuses on compressed sensing, massive MIMO, mmWave communications, NOMA, cognitive radios, DoA estimation, high-dimensional data analysis, and distributed optimization and learning.

Zhi Tian [F'13] (ztian1@gmu.edu) has been a professor with the Electrical and Computer Engineering Department of George Mason University since 2015. Prior to that, she was on the faculty of Michigan Technological University from 2000 to 2014. She served as a program director with the U.S. National Science Foundation from 2012 to 2014. Her research interests lie in the areas of signal processing, communications, detection, and estimation. Current research focuses on compressed sensing for random processes, cognitive radios and millimeter-wave communications, and decentralized network optimization. She serves on the Board of Governors of the IEEE Signal Processing Society from 2019 to 2021. She was Chair of the IEEE Signal Processing Society Big Data Special Interest Group, and a member of the IEEE Signal Processing for Communications and Networking Technical Committee. She was a Distinguished Lecturer of the IEEE Communications Society and the IEEE Vehicular Technology Society. She served as Associate Editor for *IEEE Transactions on Wireless Communications* and *IEEE Transactions on Signal Processing*. She received the IEEE Communications Society TCCN Publication Award in 2018.

Xiuzhen Cheng [F'15] (cheng@gwu.edu) is a professor with the Department of Computer Science, George Washington University, Washington, DC. She was a program director with the U.S. National Science Foundation in 2006 and from 2008 to 2010. Her research interests focus on dynamic spectrum access, cognitive radio networks, privacy-aware computing, wireless and mobile security, mobile handset networking systems (mobile health and safety), and algorithm design and analysis. She has served and is serving on the Editorial Boards of several technical journals, including *IEEE Transactions on Computers*, *IEEE Transactions on Parallel and Distributed Systems*, and *IEEE Wireless Communications*; and on the Technical Program Committees of many professional conferences/workshops, including ACM Mobihoc, ACM Mobisys, IEEE INFOCOM, IEEE ICDCS, IEEE ICC, and IEEE/ACM IWQoS. She has also chaired several international conferences, including ACM Mobihoc 2014 and IEEE PAC 2018. She is the Founder and Steering Committee Chair of the International Conference on Wireless Algorithms, Systems, and Applications, launched in 2006.