Generalized Sparse and Low-Rank Optimization for Ultra-Dense Networks

Yuanming Shi, Jun Zhang, Wei Chen, and Khaled B. Letaief

The authors present recently proposed large-scale sparse and low-rank frameworks for optimizing UDNs, supported by various motivating applications. Special attention is paid to algorithmic approaches to deal with nonconvex objective functions and constraints, as well as computational scalability.

Abstract

The ultra-dense network (UDN) is a promising technology to further evolve wireless networks and meet the diverse performance requirements of 5G networks. With abundant access points, each with communication, computation, and storage resources, the UDN brings unprecedented benefits, including significant improvement in network spectral efficiency and energy efficiency, greatly reduced latency to enable novel mobile applications, and the capability of providing massive access for Internet of Things devices. However, such great promise comes with formidable research challenges. To design and operate such complex networks with various types of resources, efficient and innovative methodologies will be needed. This motivates the recent introduction of highly structured and generalizable models for network optimization. In this article, we present some recently proposed large-scale sparse and low-rank frameworks for optimizing UDNs, supported by various motivating applications. Special attention is paid to algorithmic approaches to deal with nonconvex objective functions and constraints, as well as computational scalability.

Introduction

As mobile data traffic keeps growing at an exponential rate, and mobile applications pose more and more stringent and diverse requirements, wireless networks are facing unprecedented pressures. To further evolve wireless networks and maintain their competitiveness, network infrastructure densification stands out as a promising approach. By deploying more radio access points, supplemented with storage and computational capabilities, we can not only increase network capacity, but also improve network energy efficiency, enable low-latency mobile applications, and provide access for massive mobile devices. Such an ultra-dense network (UDN) provides an ideal platform to develop disruptive proposals to advance wireless information technologies, including cloud radio access networks (C-RANs), wireless edge caching, and mobile edge computing. These are achieved by leveraging innovative ideas in different areas, such as software-defined networking, network functions virtualization, content-centric networking, and cloud and fog computing.

By enabling capabilities of cloud computing and software-defined networking, UDNs can easily support C-RAN as an effective network architecture to exploit the benefits of network densification via centralized signal processing and interference management [1, 2]. This is achieved by moving the baseband processing functionality to the cloud data center via high-capacity fronthaul links, supported by massively deployed low-cost remote radio heads (RRHs). Meanwhile, the Internet is shifting from connection-centric to content-centric to support high-volume content delivery [3]. By enabling content caching at radio access points (i.e., *wireless edge caching*), UDNs can assist the Internet architecture evolution and achieve more efficient content delivery for mobile users [4]. Another trend is the increasing computation intensity in mobile applications, which puts a heavy burden on resource-constrained mobile devices. *Mobile edge computing* was recently proposed as a promising solution by offloading computation tasks of mobile applications to servers at nearby access points. It avoids excessive propagation delay in the backbone network compared to mobile cloud computing, and thus enables latency-critical applications. All of these systems are built on the UDN platform, which enables integration of the storage, computing, control, and networking functionalities at the ubiquitous access points. In particular, C-RANs serve the purpose of providing higher data rates, while mobile edge caching and computing networks enable low-latency content delivery and mobile applications.

However, all the emerging networking paradigms associated with UDNs bring formidable challenges to network optimization, signal processing, and resource allocation, given the highly complex network topology, the massive amount of required side information, and the high computational requirement. Typical design problems are nonconvex in nature and of enormously large scale (i.e., with large numbers of constraints and optimization variables). For example, the uncertainty or estimation error in the available channel state information (CSI) yields nonconvex quality of service (QoS) constraints, while such network performance metrics as sum throughput and energy efficiency lead to nonconvex objective functions. Thus, effective and scalable design methodologies, with the capability of handling nonconvex constraints and objectives, will be needed to fully exploit the benefits of UDNs. The aim of this article is to present recent advances in sparse and low-rank techniques for optimizing dense wireless networks [7–9], with

This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant No. 61601290, Shanghai Sailing Program under Grant No. 16YF1407700, the Hong Kong Research Grant Council under Grant No. 16200214, the National Natural Science Foundation of China under Project Nos. 61671269 and 61621091, the Chinese National 973 Program under Project No. 2013CB336600, and the 10000-Talent Program of China.

Digital Object Identifier:
10.1109/MCOM.2018.1700472

Yuanming Shi is with ShanghaiTech University; Jun Zhang is with Hong Kong University of Science and Technology; Wei Chen is with Tsinghua University; Khaled B. Letaief is with Hamad bin Khalifa University and Hong Kong University of Science and Technology.

Models	Structured sparse optimization (1)	Generalized low-rank optimization (2)
Applications	Large-scale network adaptation: 1. Network power minimization 2. User admission control 3. Active user detection	Network optimization with side information: 1. Topological interference management 2. Wireless distributed computing 3. Mobile edge caching
Algorithms	Convex optimization solver [5]: 1. $\mathcal{O}(1/k)$ convergence rate: (k: # iterations) 2. Subspace projection per iteration. Parallel cone projection per iteration	Riemannian optimization solver [6]: 1. Superlinear convergence rate with conjugate gradient 2. Quadratic convergence rate with trust region 3. Compute Riemannian gradient and Hessian per iteration

Table 1. Generalized sparse and low-rank optimization for UDNs.

comprehensive coverage including modeling, algorithm design, and theoretical analysis. We identify two representative classes of design problems in UDNs, large-scale network adaptation and side-information-assisted network optimization.

The first class of design problems are for efficient network adaptation in UDNs, including radio access point selection [7], backhaul data assignment, user admission control, user association [10], and active user detection [9]. Such large-scale network adaptation problems involve both discrete and continuous decision variables, which motivates us to enforce sparsity structures in the solutions. The success of the structured sparse optimization for network adaptation comes from the key observation that such adaptation can be achieved by enforcing structured sparsities in the solution, which are presented later in detail. The second class of design problems involve how to effectively utilize the available side information for network optimization, including topological interference management [11], wireless distributed computing [12], and mobile edge caching [4]. Network side information is critical to design UDNs, and it can take various forms, such as the network connectivity information, cache content placement at access points, and locally computed intermediate values in wireless distributed computing. We present a general incomplete matrix framework to model various network side information, which leads to a unified network performance metric via the rank of the modeling matrix for optimizing UDNs.

Although the structured sparse and low-rank techniques enjoy the benefits of modeling flexibility, the sparse function and rank function are nonconvex, which brings computational challenges [13, 14]. Furthermore, typical optimization problems in UDNs bear complicated structures, which make most of the existing algorithms and theoretical results inapplicable. To address these algorithmic challenges, we present various convexification procedures for both objectives and constraints throughout our discussion. Moreover, scalable convex optimization algorithms and nonconvex optimization techniques, such as Riemannian optimization, are presented. This article shall serve the purpose of providing network modeling methodologies and scalable computational tools for optimizing complex UDNs, as summarized in Table 1.

Structured Sparse Optimization for Large-Scale Network Adaptation

In UDNs, to effectively utilize densely deployed access points to support massive mobile devices, large-scale network adaptation will play a pivotal role. For various network adaptation problems in UDNs, the solution vector is expected to be sparse in a structured manner; for example, radio access point selection results in a group sparsity structure. To illustrate the power of the generalized sparse representation and scalable optimization paradigms, in this section, we present representative examples of group sparse beamforming for green C-RANs, and structured sparse optimization for active user detection and user admission control.

In UDNs, to effectively utilize densely deployed access points to support massive mobile devices, large-scale network adaptation will play a pivotal role. For various network adaptation problems in UDNs, the solution vector is expected to be sparse in a structured manner; for example, radio access point selection results in a group sparsity structure.

Generalized Structured Sparse Models

In this part, two motivating applications of generalized sparse models for large-scale network adaptation are presented.

Large-Scale Structured Optimization: We take green C-RAN as an example to illustrate structured optimization for network adaptation. In C-RANs, the network power consumption consists of the transmit power of active RRHs and the power of the corresponding active fronthaul links. By exploiting the spatial and temporal data traffic fluctuation, network adaptation via dynamically switching off RRHs and the associated fronthaul links can significantly reduce the network power consumption. To minimize the network power of a C-RAN, we need to optimize over both the discrete variables (i.e., the selection of RRHs and fronthaul links) and continuous variables (i.e., downlink beamforming coefficients), yielding a mixed combinatorial optimization problem, which is highly intractable. To support efficient algorithm design and analysis, a principled group sparse beamforming framework was proposed in [7] by enforcing the group sparsity structure in the solution vectors. This is achieved by a group sparsity representation of the discrete optimization variables for RRH selection as shown in Fig. 1. Specifically, by regarding all the beamforming coefficients of one RRH as a group, switching off this RRH corresponds to setting all the associated beamforming coefficients in the same group to be zeros simultaneously. We thus enforce the group sparsity structure in the aggregative beamforming vector to guide switching off the corresponding RRHs to minimize the network power consumption. Similar to group sparse beamforming for RRH selection, there is a corresponding user side node selection problem. With crowded mobile devices, it is critical to maximize the user capacity (i.e., the number of admitted users). This *user admission* problem is equivalent to minimizing the number of violated QoS constraints (modeled as $g_i(\boldsymbol{x}) \le 0$ for the ith user; it may be infeasible), which can further be modeled as minimizing the individual sparsity of the auxiliary vector $\boldsymbol{z} = [z_i]$ with $z_i \ge 0$ indicating the violations of the QoS constraints. That is, the constraint $g_i(\boldsymbol{x}) \le z_i$ (always

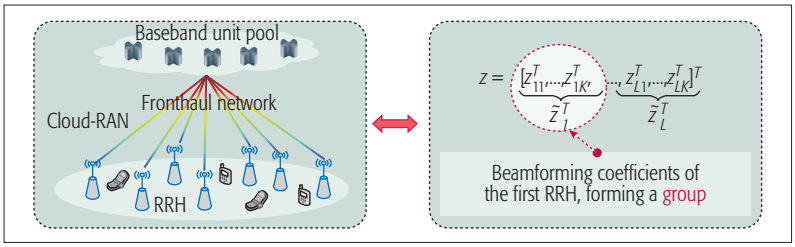

Figure 1. Group sparse beamforming for green cloud-RAN design, with L RRHs and K users. Switching off the lth RRH and the corresponding fronthaul link corresponds to setting all the beamforming coefficients in the lth beamforming coefficients group $\tilde{\boldsymbol{z}}_l = [\boldsymbol{z}_{l1}, ..., \boldsymbol{z}_{lk}$ to be zeros simultaneously, where $\boldsymbol{z}_{lk}$ is the transmit beamforming vector from RRH L to mobile user k. Note that, when switching off one RRH, the remaining RRHs need to support the QoS requirements for all the mobile users.

feasible as the auxiliary variable $z_i \geq 0$, $z_i = 0$) indicates that the original QoS constraint $g_i(\boldsymbol{x}) \leq 0$ is feasible, while $z_i > 0$ indicates that the original QoS constraint $g_i(\boldsymbol{x}) \leq 0$ is infeasible. Therefore, by enforcing this structured sparsity in the solution, user admission can be effectively handled.

High-Dimensional Structured Estimation: With limited radio resources, it is challenging to support massive device connectivity for such applications as IoT. Fortunately, only part of the massive devices will be active at a time given the sporadic traffic for the emerging applications (e.g., machine-type communications, Internet of Things [IoT]) [9]. Active user detection is thus a key problem for providing massive connectivity in UDNs, which turns out to be a structured sparse estimation problem. Specifically, suppose we have N single-antenna mobile devices (K of which are active) and one M-antenna base station (BS). The received signal at the BS has the form $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{\Sigma}\boldsymbol{Q} + \boldsymbol{W}$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ is the unknown diagonal activity matrix with K non-zero diagonals whose positions are to be estimated, $\boldsymbol{H} \in \mathbb{C}^{M\times N}$ is the unknown channel matrix from all the devices to the BS, $\boldsymbol{Q} \in \mathbb{C}^{N\times L}$ is the known pilot matrix with training length L, and $\boldsymbol{W}$ is the additive noise. We thus need to simultaneously estimate the channel matrix $\boldsymbol{H}$ and $\boldsymbol{\Sigma}$, which poses a great challenge. We observe that detecting the active users is equivalent to estimating the group sparsity structure of the combined matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma} \in \mathbb{C}^{M\times N}$, which has a group structured sparsity in columns of matrix $\boldsymbol{\Theta}$ induced by the structure of $\boldsymbol{\Sigma}$. That is, when mobile device N is inactive, all the entries in the Nth column in matrix $\boldsymbol{\Theta}$ become zeros simultaneously. Due to the limited radio resources, the training length L will be much smaller than the channel dimension N, and thus, the estimation problem is ill-posed and yields a high-dimensional structured estimation problem.

Fortunately, the embedded low-dimensional structure (i.e., the structured sparsity) can be algorithmically exploited to ensure the success of the high-dimensional structured estimation, as illustrated in Fig. 2 for the behaviors of phase transitions and normalized mean square error (NMSE) . Phase transition defines a sharp change in the behavior of a computational problem as its parameters vary. Convex geometry and conic integral geometry provide principled ways to theoretically predicate the phase transitions precisely [15]. In particular, the phase transition phenomenon in Fig. 2a reveals the fundamental limits of sparsity recovery in the best cases (i.e., without noise). Specifically, such study reveals that the required training length, or the number of measurements, depends on the sparsity level of $\boldsymbol{\Theta}$, and highly accurate user activity detection can be achieved with sufficient measurements. Figure 2b further demonstrates that the low-dimensional structure can be exploited to significantly reduce the training length for active user detection even in noisy scenarios.

A Generalized Sparse Optimization Paradigm

We have demonstrated that effective network adaptation can be achieved by either inducing vector sparsity in the structured manner or estimating the structured sparsity pattern. In this part, we provide a generalized sparse optimization framework to algorithmically exploit the low-dimensional structures in UDNs. This is achieved by optimizing a constrained composite combinatorial objective,

$$\underset{\mathbf{z}\in\mathbb{C}^n}{\text{minimize}}\ f(\mathbf{z}) \coloneqq f_1(Supp(\mathbf{z})) + f_2(\mathbf{z}) \quad \text{subject to } \mathbf{z} \in \mathcal{C}, \tag{1}$$

where Supp($\boldsymbol{z}$) is the index set of non-zero coefficients of a vector $\boldsymbol{z}$, f_1 is a combinatorial positive-valued set-function to control the structured sparsity in $\boldsymbol{z}$, f_2 is a continuous convex function in $\boldsymbol{z}$ to represent the system performance such as transmit power consumption, and the constraint set $\mathcal{C}$ serves the purpose of modeling system constraints (e.g, transmit power constraints and QoS constraints). The most natural convex surrogate for a nonconvex function f is its convex envelope (i.e., its tightest convex lower bound). The main motivation for convexifying function f is that the convexified optimization problems make it possible to use the convex geometry theory [15] to reveal benign properties about the globally optimal solutions, which can be computed with efficient algorithms. For example, the individual sparsity function with ℓ_0-norm in $\boldsymbol{z}$ can be convexified to the ℓ_1-norm. The group sparsity function can be convexified by the mixed ℓ_1/ℓ_2-norm. More general convex relaxation results can be derived based on the principles of convex analysis [15]. Note that it is critical to establish the optimality for various convex relaxation approaches in UDNs. For example, for the nonconvex active user detection problem from earlier, the optimality condition can be established via the conic geometry approach in [15].

The constraint set $\mathcal{C}$ serves the purpose of modeling various QoS constraints including unicast beamforming, multicast beamforming, and stochastic beamforming, just to name a few. For example, the nonconvex QoS constraints for unicast beamforming can be equivalently transformed into convex second-order cone constraints [7]. Furthermore, physical layer integration techniques can effectively improve the network performance via providing multicast services, which, however, yield nonconvex quadratic QoS constraints. The semidefinite relaxation (SDR) technique turns out to be effective to convexify the nonconvex quadratic constraints via lifting the original vector problem to higher matrix dimensions, followed by dropping the rank-one constraints. For stochastic beamforming with probabilistic QoS constraints due to CSI uncertainty, the probabilistic QoS constraints can be convexified based on the principles of the majorization-minimization procedure, yielding sequential convex approximations. In

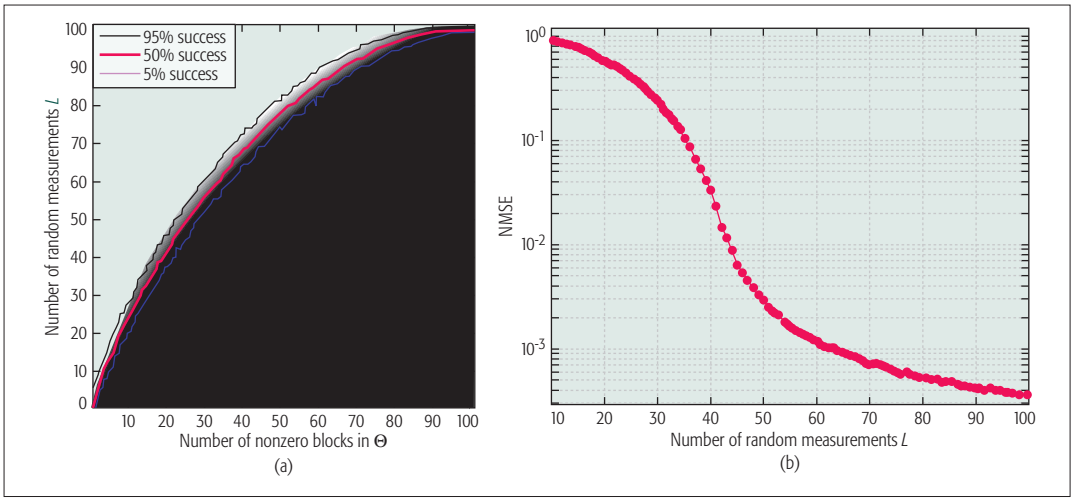

Figure 2. a) Phase transitions for structured sparse estimation for massive device connectivity, given the random measurement $Y = \Theta Q$ with $N = 100$ and $M = 2$. The heat map indicates the empirical probability of success (black = 0%; white = 100%); b) NMSE of estimating Θ, given the noisy model $\boldsymbol{Y} = \Theta \boldsymbol{Q} + \boldsymbol{W}$ with $N = 100$, $M = 2$, $K = 20$, and $W_{ij} \sim \mathcal{CN}(0,0.01)$. Each entry in $\boldsymbol{Q} \in \mathcal{C}^{N \times L}$ is distributed as $Q_{ij} \sim \mathcal{CN}(0,1)$ with L as the number of random measurements.

summary, the general formulation in Eq. 1 enables efficient algorithm design and analysis for network adaptation in UDNs.

Generalized Low-Rank Optimization with Network Side Information

UDNs are highly complex to optimize, for which it is critical to exploit the available network side information. For example, network connectivity information, cached content at the access points, and locally computed intermediate values all serve as exploitable side information for efficiently designing coding and decoding in UDNs.

In this section, we provide a generalized low-rank matrix modeling framework to exploit the network side information, which helps to efficiently optimize across the communication, computation, and storage resources. To demonstrate the power of this framework, we present topological interference alignment as a concrete example and then extend it to cache-aided interference channels and wireless distributed computing systems. A general low-rank optimization problem is then formulated by incorporating the network side information.

Network Side Information Modeling via Incomplete Matrix

To exploit the full performance gains of network densification, recent years have seen progress on interference management under various scenarios depending on the amount of shared CSI and user messages. Typical interference management strategies include interference alignment, interference coordination, and coordinated multipoint transmission and reception, to name just a few. However, the significant overhead of acquiring global CSI motivates numerous research efforts on CSI overhead reduction strategies (e.g., delayed CSI, alternating CSI, and mixed CSI). One of the most promising strategies is topological interference management (TIM) [11], for which only network connectivity information is required. This is based on the fact that most of the wireless channel propagation links are weak enough to be ignored, thanks to path loss and shadowing. However, the TIM problem turns out to be linear index coding problems [11], which are in general highly intractable, and only partial results exist for special cases. Recently, a new proposal was made for the TIM problem, which can greatly assist the algorithm design. The main innovation is to model the network connectivity pattern in UDNs as an incomplete matrix. Then the TIM problem can be formulated as a generalized matrix completion problem,[1] which helps to develop effective linear precoding and decoding strategies. Figure 3 demonstrates the modeling framework, with Fig. 3a showing a five-user interference channel as an example and Fig. 3c showing the corresponding modeling matrix. The task of TIM is to complete the side information modeling matrix, which will then determine the precoder and decoder [8].

This modeling framework is very powerful, and can be adopted to consider other design problems in UDNs. By equipping the densely deployed radio access points and mobile devices with isolated cache storages, caching the content at the edge of the network provides a promising way to improve the throughput and reduce latency, as well as reduce the load of the core network and RANs [4]. In general, content-centric communications consist of two phases, a content placement phase followed by a content delivery phase. However, due to the coupled wireline and wireless communications in cache-aided UDNs, unique challenges arise in the edge caching problem. Fortunately, the incomplete matrix modeling framework can capture the information of the content cached at different nodes. Figure 3b shows an example for cache-aided five-user interference channels, where the side information is represented in the side information modeling matrix in Fig. 3c. Similarly, this modeling framework can also be extended to wireless distributed computing networks [12]. For the prevalent distributed computing structures like MapReduce and Spark, the basic idea is that intermediate values computed in the "map" phase based on the locally available dataset can be regarded as the side information for the "reduce" phase to compute the output value

UDNs are highly complex to optimize, for which it is critical to exploit the available network side information. For example, network connectivity information, cached content at the access points, and locally computed intermediate values, all serve as exploitable side information for efficiently designing coding and decoding in UDNs.

[1] B. Hassibi, "Topological Interference Alignment in Wireless Networks," *Smart Antennas Workshop*, Aug. 2014.

The cubic computational complexity of each Newton step limits its capability to scale to large network sizes in UDNs. This motivates enormous research efforts to improve the computational efficiency for convex programs, including the techniques of first-order methods, randomization, parallel and distributed computing.

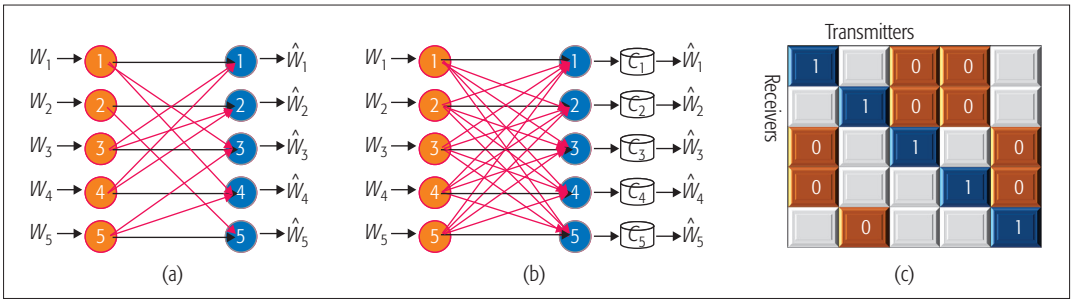

Figure 3. a) A partially connected 5-user interference channel with the index set of connected links as $\mathcal{V}$ = {(1, 3), (1, 4), (2, 3), (2, 4), (3, 1), (3, 5), (4, 1), (4, 5), (5, 1)}; b) a cache-aided 5-user interference channel with cached messages at each receiver indexed by $\mathcal{C}_1 = \{2, 5\}$, $\mathcal{C}_2 = \{1,5\}$, $\mathcal{C}_3 = \{2, 4\}$, $\mathcal{C}_4 = \{2, 3\}$, and $\mathcal{C}_5 = \{1, 3, 4\}$; c) let $\boldsymbol{M} = [M_{ij}] \in \mathbb{C}^{K\times K}$ with $M_{ij} = \boldsymbol{u}_i^{\mathrm{H}}\boldsymbol{v}_i \in \mathbb{C}$, where $\boldsymbol{u}_i \in \mathbb{C}^n$ and $\boldsymbol{v}_i \in \mathbb{C}^n$ are the precoding and decoding vectors with N as the number of channel uses for transmission. The incomplete matrix $\boldsymbol{M}$ with partial known entries indexed by $\mathcal{V}$ serves as the side information modeling matrix for a) and b); that is, $M_{ii} = 1$ means to preserve the desired signals for each receiver, $M_{ij} = 0$, $\forall (i, j) \in \mathcal{V}$ represents cancelling the interference, and $M_{ij} = *$, $\forall (i, j) \notin \mathcal{V} \cup \{(i, i)\}$ can be any (unknown) values.

for a given input. This can then help reduce the communication overhead in the "shuffle" phase to obtain the intermediate values that are not computed locally in the "map" phase. The incomplete matrix modeling approach will help to formulate the design problems for wireless caching and distributed computing systems.

A Generalized Low-Rank Optimization Paradigm

We have presented an effective and general framework to model various network side information in UDNs. Next, we present a low-rank optimization formulation to exploit the available network side information. The side information modeling matrix $\boldsymbol{M}$ as shown in Fig. 3c helps cancel interference over N channel uses, yielding an interference-free channel with $1/N$ degrees of freedom (DoF) (i.e., the first-order data characterization). Observe that the rank of the side information modeling matrix $\boldsymbol{M}$, denoted by rank ($\boldsymbol{M}$), equals the number of channel uses N, which equals the inverse of the achievable DoF. To maximize the achievable DoF, we thus can minimize the rank of the side information modeling matrix, yielding the following generalized low-rank optimization problem,

$$\underset{\mathbf{M}\in\mathbb{C}^{p\times q}}{\text{minimize}}\ \text{rank}(\mathbf{M})\ \ \text{subject to}\ \mathbf{M}\in\mathcal{D}, \tag{2}$$

where the constraint set $\mathcal{D}$ encodes the network side information. Low-rank optimization has been proved to be a key design tool in machine learning, high-dimensional statistics, signal processing, and computational mathematics [14]. The rank function is nonconvex and thus is computationally difficult, but convexifying it leads to efficient algorithms. For example, the nuclear norm (i.e., the summation of singular valves of a matrix) provides a convex surrogate of the rank function that is analogous to the ℓ_1-norm relaxation of the cardinality of a vector.

Given the special structure of the side information modeling matrix in UDNs, most existing algorithmic and theoretical results for low-rank optimization are inapplicable. Recent work [8] contributed a novel proposal of nonconvex paradigms for solving the generalized low-rank optimization problem (Eq. 2) by optimizing over the nonconvex rank constraints directly via Riemannian optimization and matrix factorization. Figure 4 illustrates the phase transition behavior for the generalized low-rank optimization in topological interference management, which characterizes the relationships between the achievable DoF and the number of connected interference links on average. Given the rank, representing the achievable DoF, with more connected interference links, the success probability for recovering the incomplete side information modeling matrix is lower. It thus provides the guidelines for network deployment in dense wireless networks, content placement in cache-aided interference channels, and dataset placement in wireless distributed computing systems.

Optimization Algorithms and Analysis

We have seen quite a few algorithmic challenges for the sparse and low-rank modeling frameworks for UDNs. In this section, we present some new trends in optimization algorithms for solving the generalized sparse and low-rank optimization problems in the forms of Eqs. 1 and 2, respectively. Basically, numerical optimization algorithms can be classified in terms of first vs. second order methods, depending on whether they use only gradient-based information vs. calculations of both the first and second derivatives. The convergence rates of second-order methods are usually faster with the caveat that each iteration is more expensive. In general, there is a trade-off between the per-iteration computation cost vs. the total number of iterations, although first-order methods often scale better to large-scale high-dimensional statistics problems [13]. While optimization problems in communication systems are typically solved in the convex paradigm with the second-order methods, thanks to the ease of use of the CVX toolbox, we have observed the necessity of the first-order methods and the importance of the nonconvex paradigm, as elaborated in the following subsections.

Convex Optimization Algorithms

We have presented a variety of methodologies to convexify the nonconvex objective functions and nonconvex constraints for the generalized sparse optimization problem (Eq. 1). Newton iteration-based interior-point methods supported by many user-friendly software packages (e.g., CVX) provide a general way to solve constrained convex optimization problems. However, the

cubic computational complexity of each Newton step limits its capability to scale to large network sizes in UDNs. This motivates enormous research efforts to improve the computational efficiency for convex programs, including the techniques of first-order methods, randomization, and parallel and distributed computing.

Parallel and distributed optimization provides a principled way to exploit the distributed computing environments to increase the levels of scalability while reducing the communication costs. To solve a general large-scale convex program, a principled two-stage framework has recently been proposed in [5] with the capability of providing certificates of infeasibility, enabling parallel and scale computing. This is achieved, in the first stage, by the matrix stuffing technique to quickly transform the original convex programs into the standard conic optimization problem form via updating the associated values in the pre-stored structure of the standard conic program. In the second stage, the ADMM-based algorithm is adopted to solve the standard large-scale conic optimization problem via exploiting the problem structures [5] to enable parallel cone projection at each iteration.

Other lines of work have focused on the use of first-order methods and randomization to solve large convex programs. In particular, for sparse convex optimization problems, Frank-Wolfe-type algorithms (a.k.a. conditional gradient) have recently gained enormous interest, fueled by the excellent scalability with projection-free operations via exploiting the well structured sparsity constraints. The coordinate descent method has gained popularity for scalability by choosing a single coordinate (or a block of coordinates) to be updated within each iteration, thereby reducing the iteration computing cost. Approximation techniques, including randomization methods and sketching methods, further provide algorithmic opportunities to enable scalability for, in particular, first-order methods, via speeding up numerical linear algebra or reducing problem dimensions. In particular, the stochastic gradient method provides a generic way to stochastically approximate the gradient descent method to solve large-scale machine learning problems. All the above presented algorithmic and theoretical results may be leveraged to solve large-scale convex optimization problems in UDNs.

Nonconvex Optimization Algorithms

Recently, a new line of work has attracted significant attention, which focuses on solving the nonconvex optimization problems directly via developing efficient nonconvex procedures, sometimes with optimality guarantee. We have seen recent progress on nonconvex procedures based on various algorithms (e.g., projected/stochastic/conditional gradient methods, Riemannian manifold optimization algorithms) for a class of high-dimensional statistical problems and machine learning problems, including low-rank matrix completion, phase retrieval, and blind deconvolution, to name just a few. In particular, optimization by directly exploiting problems' manifold structures is becoming a general and powerful approach to solve various nonconvex optimization problems. The structured constraints such as rank and orthogonality appear in many machine learning applications, including sensor network localization, dimensionality reduction, low-rank matrix recovery, phase synchronization, and community detection.

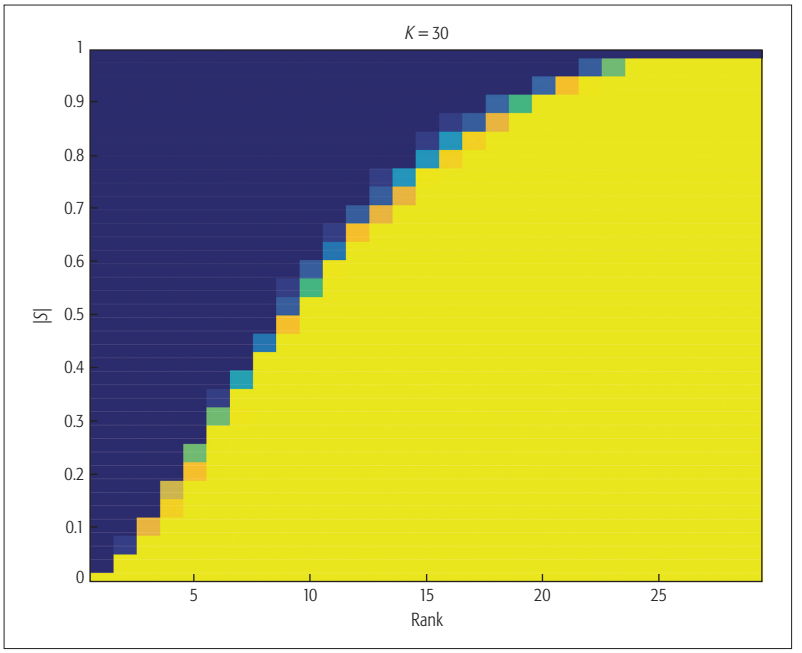

Figure 4. Phase transitions for the topological interference management problem for a partially connected K-user interference channel with the network side information modeling constraint set $\mathcal{D} = \{M \in \mathbb{R}^{30\times 30} \,|\, M_{ii} = 1,\ M_{ij} = 0,\ \forall (i, j) \in \mathcal{S}\}$, where the set $\mathcal{S}$ is randomly and uniformly sampled. The heat map indicates the empirical probability of success (blue = 0%; yellow = 100%).

From a high-level standpoint, Riemannian optimization is the extension of standard unconstrained optimization searching in the Euclidean space to optimization in the Riemannian manifold space by generalizing concepts such as the gradient and Hessian [6]. A graphic representation of Riemannian optimization algorithms is illustrated in Fig. 5. Specifically, the Euclidean gradient $\nabla f(\boldsymbol{X}_k)$ needs to be projected to the tangent space $T_{\boldsymbol{X}_k}\mathcal{M}$ of manifold $\mathcal{M}$ to define a search direction $\xi_{\boldsymbol{X}_k}$ (which can be computed based on the principles of the conjugate gradient method or trust-region method), followed by the retraction operator $\mathcal{R}_{\boldsymbol{X}_k}$ to define a new iteration $\boldsymbol{x}_{k+1} = \mathcal{R}_{\boldsymbol{X}_k}(\alpha\xi_k)$ (α is the step size) on the manifold $\mathcal{M}$. In particular, we exploit the manifold geometry of fixed-rank matrices to solve the low-rank optimization problem (Eq. 2) efficiently. Figure 5b demonstrates the effectiveness of Riemannian optimization-based methods. It shows that the Riemannian optimization enjoys fast convergence rates, for example, compared to an existing approach based on alternating minimization.

Conclusions and Future Directions

This article presents generalized sparse and low-rank optimization techniques for optimizing across communication, computation, and storage resources in UDNs by exploiting network structures and side information. Illustrated by important application examples, various structured sparse modeling methods are introduced, and an incomplete matrix representation is presented to model different types of network side information. Methodologies of designing scalable algorithms are discussed, including both convex and nonconvex methods. The presented results and methodologies demon-

Generalized sparse and low-rank optimization techniques are mainly applied to improve the network energy efficiency and spectral efficiency in UDNs. However, emerging mobile applications have strong demands for user privacy and ultra-low latency communications, which call for more general mathematical models and formulations.

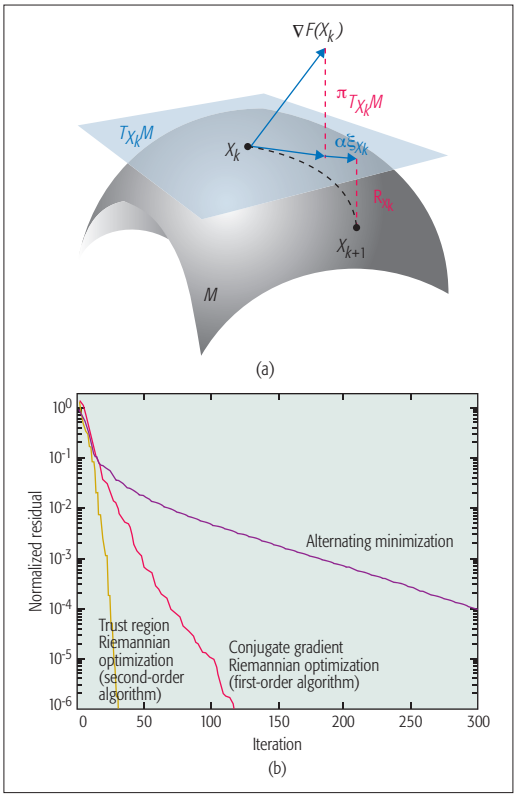

Figure 5. a) Graphic representation of Riemannian optimization algorithms at each iteration; b) solving rank-constrained optimization problem minimize$_{M\in\mathcal{M}}$ $f(M)$ with $\mathcal{M} = \{\boldsymbol{M} \in \mathbb{R}^{100\times100} \,|\, \text{rank}(M) = 5\}$ and $f(M) = \Sigma_{i=1}^{100}(M_{ii} - 1)^2 + \Sigma_{(i,j)\in\Omega}M_{ij}^2$ ($|\Omega| = 400$) via Riemannian optimization algorithms by factorizing rank-r matrix M, yielding a quotient manifold $[M] = \{(\boldsymbol{UQ}_U, \boldsymbol{Q}_U^T\Sigma\boldsymbol{Q}_V, \boldsymbol{VQ}_V)$: $\boldsymbol{Q}_U, \boldsymbol{Q}_V \in \mathcal{Q}(r)\}$, where $\mathcal{O}(r)$ is the set of all $r \times r$ orthogonal matrices.

strate the effectiveness of structured optimization techniques for designing UDNs.

Despite the encouraging progress, there still remain a variety of interesting open questions. To date, generalized sparse and low-rank optimization techniques are mainly applied to improve the network energy efficiency and spectral efficiency in UDNs. However, emerging mobile applications have strong demands for user privacy and ultra-low-latency communications, which call for more general mathematical models and formulations. Other interesting problems concern the theoretical analysis for the generalized sparse and low-rank optimization models and algorithms. Although we have seen significant progress in theoretical understanding of sparse and low-rank optimization problems via convex relaxation approaches [15] and nonconvex procedures, it is challenging to apply existing results to the generalized sparse and low-rank optimization problems (Eqs. 1 and 2) due to the complicated structures. Finally, there are a variety of interesting research directions associated with improving the computational scaling behavior of various algorithms via recent proposals (e.g., randomized algorithms based on sketching).

References

[1] T. Q. S. Quek *et al.*, *Cloud Radio Access Networks: Principles, Technologies, and Applications*, Cambridge Univ. Press, 2017.

[2] H. Zhang *et al.*, "Fronthauling for 5G LTE-U Ultra Dense Cloud Small Cell Networks," *IEEE Wireless Commun.*, vol. 23, no. 6, Dec. 2016, pp. 48–53.

[3] G. Xylomenos *et al.*, "A Survey of Information-Centric Networking Research," *IEEE Commun. Surveys & Tutorials*, vol. 16, vol. 2, 2014, pp. 1024–49.

[4] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 82–89.

[5] Y. Shi *et al.*, "Large-Scale Convex Optimization for Dense Wireless Cooperative Networks," *IEEE Trans. Signal Processing*, vol. 63, Sept. 2015, pp. 4729–43.

[6] N. Boumal *et al.*, "Manopt, A Matlab Toolbox for Optimization on Manifolds," *J. Mach. Learn. Res.*, vol. 15, 2014, pp. 1455–59.

[7] Y. Shi, J. Zhang, and K. B. Letaief, "Group Sparse Beamforming for Green Cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, May 2014, pp. 2809–23.

[8] Y. Shi, J. Zhang, and K. B. Letaief, "Low-Rank Matrix Completion for Topological Interference Management by Riemannian Pursuit," *IEEE Trans. Wireless Commun.*, vol. 15, July 2016, pp. 4703–17.

[9] G. Wunder *et al.*, "Sparse Signal Processing Concepts for Efficient 5G System Design," *IEEE Access*, vol. 3, 2015, pp. 195–208.

[10] H. Zhang *et al.*, "Energy Efficient User Association and Power Allocation in Millimeter-Wave-Based Ultra Dense Networks with Energy Harvesting Base Stations," *IEEE JSAC*, vol. 35, Sept. 2017, pp. 1936–47.

[11] S. Jafar, "Topological Interference Management Through Index Coding," *IEEE Trans. Info. Theory*, vol. 60, Jan. 2014, pp. 529–68.

[12] S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "Coding for Distributed Fog Computing," *IEEE Commun. Mag.*, vol. 55, no. 4, Apr. 2017, pp. 34–40.

[13] M. J. Wainwright, "Structured Regularizers for High-Dimensional Problems: Statistical and Computational Issues," *Annual Rev. Stat. Appl.*, vol. 1, 2014, pp. 233–53.

[14] M. A. Davenport and J. Romberg, "An Overview of Low-Rank Matrix Recovery from Incomplete Observations," *IEEE J. Selected Topics in Signal Processing*, vol. 10, June 2016, pp. 608–22.

[15] D. Amelunxen *et al.*, "Living on the Edge: Phase Transitions in Convex Programs with Random Data," *Info. Inference*, vol. 3, June 2014, pp. 224–94.

Biographies

Yuanming Shi [S'13, M'15] (shiym@shanghaitech.edu.cn) received his B.S. degree from Tsinghua University in 2011 and his Ph.D. degree from Hong Kong University of Science and Technology (HKUST) in 2015. He is currently an assistant professor at ShanghaiTech University. He received the 2016 IEEE Marconi Prize Paper Award and the 2016 Young Author Best Paper Award from the IEEE Signal Processing Society. His research interests include dense wireless networks, intelligent IoT, mobile AI, machine learning, statistics, and optimization.

Jun Zhang [M'10, SM'15] (eejzhang@ust.hk) received his Ph.D. degree from the University of Texas at Austin. He is currently a research assistant professor at HKUST. He received the 2016 Marconi Prize Paper Award in Wireless Communications and the 2016 IEEE ComSoc Asia-Pacific Best Young Researcher Award. His research interests include dense wireless cooperative networks, mobile edge caching and computing, cloud computing, and big data analytics systems.

Wei Chen [S'05, M'07, SM'13] (wchen@tsinghua.edu.cn) received his B.S. and Ph.D. degrees (Hons.) from Tsinghua University in 2002 and 2007, respectively. Since 2007, he has been on the faculty at Tsinghua University, where he is a tenured full professor and a member of the University Council. He is a member of the National 10,000-Talent Program and a Cheung Kong Young Scholar. He received the IEEE Marconi Prize Paper Award and the IEEE Comsoc Asia Pacific Board Best Young Researcher Award.

Khaled B. Letaief [S'85, M'86, SM'97, F'03] (eekhaled@ust.hk) received his Ph.D. degree from Purdue University. From 1990 to 1993, he was a faculty member at the University of Melbourne, Australia. He has been with HKUST since 1993 where he was Dean of Engineering. In September 2015, he joined HBKU in Qatar as Provost. He is an ISI Highly Cited Researcher and a recipient of many distinguished awards. He has served in many IEEE leadership positions including ComSoc President (at present), Vice-President for Technical Activities, and Vice-President for Conferences.