A Random Access Protocol for Pilot Allocation in Crowded Massive MIMO Systems

Emil Björnson, *Member, IEEE*, Elisabeth de Carvalho, *Senior Member, IEEE*, Jesper H. Sørensen, *Member, IEEE*, Erik G. Larsson, *Fellow, IEEE*, Petar Popovski, *Fellow, IEEE*

Abstract—The massive multiple-input multiple-output (MIMO) technology has great potential to manage the rapid growth of wireless data traffic. Massive MIMO achieves tremendous spectral efficiency by spatial multiplexing many tens of user equipments (UEs). These gains are only achieved in practice if many more UEs can connect efficiently to the network than today. As the number of UEs increases, while each UE intermittently accesses the network, the random access functionality becomes essential to share the limited number of pilots among the UEs. In this paper, we revisit the random access problem in the Massive MIMO context and develop a reengineered protocol, termed *strongest-user collision resolution* (SUCRe). An accessing UE asks for a dedicated pilot by sending an uncoordinated random access pilot, with a risk that other UEs send the same pilot. The favorable propagation of massive MIMO channels is utilized to enable distributed collision detection at each UE, thereby determining the strength of the contenders' signals and deciding to repeat the pilot if the UE judges that its signal at the receiver is the strongest. The SUCRe protocol resolves the vast majority of all pilot collisions in crowded urban scenarios and continues to admit UEs efficiently in overloaded networks.

Index Terms—Massive MIMO, random access, collision resolution.

I. Introduction

THE number of wirelessly connected devices and their respective data traffic are growing rapidly as we transition into the fully networked society [1]. The growth is currently driven by video streaming, social networking, and new use cases, such as machine-to-machine communication and internet of things. In dense urban scenarios, the METIS project predicts a future with up to 200.000 devices per km^2 and an associated data volume of 500 Gbyte/month/device [2]. These are massive numbers that call for radical changes in

E. Björnson and E. G. Larsson are with the Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden (e-mail: emil.bjornson@liu.se; erik.g.larsson@liu.se).

E. de Carvalho, J. H. Sørensen, and P. Popovski are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg Øst, Denmark (e-mail: edc@es.aau.dk; petarp@es.aau.dk).

the network infrastructure to avoid congestion and guarantee service quality and availability. A fair amount of the traffic is expected to be offloaded to WiFi and small-cell-technology operating at mm-wave frequencies, but macro cellular networks operating at frequencies of one or a few GHz will remain to define the coverage, mobility support, and guaranteed service quality. Hence, future cellular networks need to handle urban deployment with *massive* numbers of connected UEs that request *massive* data volumes [3].

Since the cellular frequency bands are scarce, orders-of-magnitude capacity improvements are only possible by radically increasing the spectral efficiency (SE) [bit/s/Hz]. The Massive MIMO network topology, proposed in the seminal paper [4], can theoretically deliver such extraordinary improvements [5]. The gains are achieved by spatial multiplexing, where base stations (BSs) with hundreds of antennas are utilized to serve tens of UEs per cell, at the same time-frequency resource. Massive MIMO is primarily a technology for time-division duplexing (TDD) [6], where scalable channel estimation protocols are achieved by only requiring uplink pilots and utilizing channel reciprocity to obtain downlink channel estimates [7]. The communication-theoretic performance and asymptotic limits of Massive MIMO have been analyzed extensively in recent years; see for instance [4]–[6], [8]–[15]. These works have focused on the performance achieved by *active* UEs, mainly in homogeneously loaded cells, while the network access functionality has received little attention [16]. It is important to note that the building block of the traffic is data packets, which are generated in an intermittent and unpredictable fashion in all non-streaming applications. For example, web browsing and social network applications are characterized by bursty on-off traffic where intensive download activities are interwoven with long periods of silence. Many UEs thus switch between being active and inactive at a frequent and irregular basis. Since Massive MIMO will be deployed to handle, say, 10× higher UE densities than in current systems, the number of UEs that switch between active and inactive mode in a given time interval will be 10× higher as well. Scalable and efficient access protocols are thus mandatory.

A. Random Access Functionality in LTE

Before we propose a new highly scalable access protocol, we describe the conventional protocol used on the Physical Random Access Channel (PRACH) of Long-Term Evolution (LTE). It consists of four steps [17], as illustrated in Fig. 1.
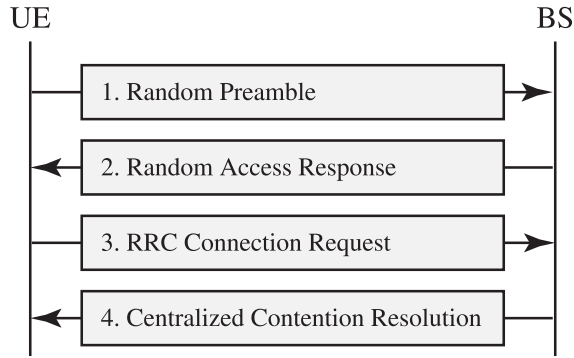
Fig. 1. PRACH protocol of the LTE system.

In Step 1, each accessing UE picks a preamble at random from a predefined set. The preamble is an entity that enables synchronization towards the BS. It does not carry specific reservation information or data and thus can be viewed as a pilot sequence. Since UEs that wish to access the network are not coordinated in picking the preambles, a collision occurs if two or more UEs select the same preamble simultaneously. A BS in LTE only detects if a specific preamble is active or not in Step 1 [18]. In Step 2, the BS sends a *random access response* corresponding to each activated preamble, to convey physical parameters (e.g., timing advance) and allocate a resource to the UE (or UEs) that activated the preamble. In Step 3, each UE that has received a response to its preamble transmission sends a *RRC (Radio Resource Control) Connection Request* in order to request resources for subsequent data transmission. If more than one UE activated the preamble, then all these UEs use the same resource to send their RRC connection request in Step 3 and this collision is detected by the BS. Step 4 is called *contention resolution* and contains one or multiple steps to resolve the collision. This is a complicated procedure that might result in that all colliding UEs need to make a new access attempt after a random waiting time.

B. Prior Work on Random Access in Massive MIMO

Conventional cellular networks allocate dedicated resources to each active UE, thus the BS needs to convey the time-frequency positions of these resources. In contrast, Massive MIMO systems allocate all time-frequency resources to all active UEs, and separate them spatially based on their pilot sequences. The number of pilots is limited by the size of the channel coherence block. In the original Massive MIMO concept of [4] the UEs within a cell use mutually orthogonal pilots, while the necessary reuse of pilots across cells causes *inter-cell pilot contamination* that leads to additional interference [8].

In crowded urban scenarios, the total number of active and inactive UEs residing in a cell is also much larger than the number of available pilot sequences. Hence, a pilot cannot be permanently associated with a UE in the cell but need to be opportunistically allocated and deallocated to follow its unpredictable intermittent data traffic pattern. Random access (RA) mechanisms can be used for such pilot allocation, but must be designed to resolve collisions in crowded

cells without excessive access delays. The papers [19]–[21] propose protocols where UEs can transmit data whenever they like by selecting pilots at random from a common pool. This eliminates access delays, at the cost of pilot collisions that cause *intra-cell pilot contamination*. The collisions are expressed as a graph code in [19] and belief propagation is used to mitigate pilot contamination, by utilizing the principles of coded random access [22], [23]. Several SE expressions are derived in [20] and utilized to optimize the UE activation probability and pilot length. Asynchronous UE timings are utilized in [21] to detect and resolve pilot collisions in RA.

In this paper, we explore the alternative solution that each UE needs to be assigned a pilot sequence before transmitting payload data, to avoid intra-cell pilot collisions and actualize the payload transmission situation considered in the main body of Massive MIMO research [4]–[6], [8]–[11]. We focus on urban deployments with small initial timing variations and propose a new RA protocol for UEs that wish to access the network. The protocol can resolve RA collisions in a distributed and scalable way, by exploiting special properties of Massive MIMO channels. A preliminary version of the protocol was described in [24].

C. Outline and Notation

The remainder of this paper is organized as follows. Section II describes the proposed RA protocol assuming arbitrary Massive MIMO channels, while Section III analyzes the performance with uncorrelated Rayleigh fading channels. Section IV provides numerical results that show how the proposed protocol can resolve severe RA collisions and operate under very high load. The main conclusions are summarized in Section V.

The following notation is used throughout the paper. The transpose, conjugate-transpose, and conjugate of a matrix $\mathbf{X}$ are denoted by $\mathbf{X}^{\mathrm{T}}$, $\mathbf{X}^{\mathrm{H}}$, and $\mathbf{X}^*$, respectively. We let $\mathbf{I}_M$ denote the $M \times M$ identity matrix, whereas $\| \cdot \|$ and $| \cdot |$ stand for the Euclidean norm of a vector and cardinality of a set, respectively. The notations $\mathbb{E}\{\cdot\}$ and $\mathbb{V}\{\cdot\}$ indicate the expectation and variance with respect to a random variable. We use $\mathcal{N}(\cdot, \cdot)$ to denote a Gaussian distribution, $\mathcal{CN}(\cdot, \cdot)$ to denote a circularly-symmetric complex Gaussian distribution, $B(\cdot, \cdot)$ to denote a binomial distribution, and χ_n to denote a chi-distribution with n degrees of freedom. We use $\mathbb{C}$ and $\mathbb{R}$ to denote spaces of complex-valued and real-valued numbers, respectively. The Gamma function is denoted by $\Gamma(\cdot)$.

II. PROPOSED RANDOM ACCESS PROTOCOL

We consider cellular networks where each BS is equipped with M antennas. The system operates in TDD mode and the time-frequency resources are divided into coherence blocks of T channel uses, dimensioned such that the channel responses between each BS and its UEs are constant and frequency flat within a block, while they vary between blocks. This can be implemented using orthogonal frequency-division multiplexing (OFDM). We let $\mathcal{U}_i$ denote the set of UEs that reside in cell i. At any given time, only a subset $\mathcal{A}_i \subset \mathcal{U}_i$ of the UEs are active in the sense that they are transmitting and/or receiving data. Note that in the scenarios relevant for Massive
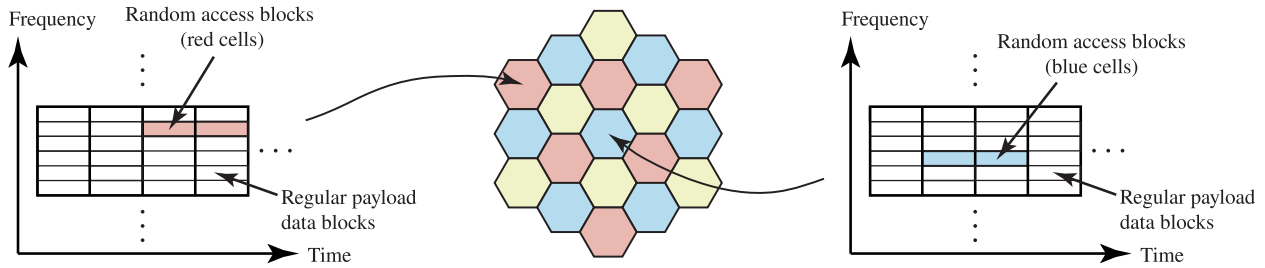
Fig. 2. Illustration of the proposed transmission protocol, where the time-frequency domain is divided into coherence blocks. The majority of the blocks are used for payload data transmission for active UEs (which have been allocated dedicated pilots). A few of the blocks are used for random access where inactive UEs can ask to be admitted to the payload data blocks. The location of the random access blocks in the time-frequency grid is different between adjacent cells (here illustrated using a reuse factor of three), and the neighbors are either quiet or send payload during the RA.

MIMO deployment, we typically have a very large UE set: $|\mathcal{U}_i| \gg T$. However, the active UEs satisfy $|\mathcal{A}_i| < T$, thus the BS can temporarily assign orthogonal pilot sequences to these UEs and reclaim them when their respective transmissions are finished.

The coherence blocks are divided into two categories: *payload data blocks* and *random access blocks*. For each cell i, the first category is used for uplink (UL) and downlink (DL) data transmission to the UEs in the set $\mathcal{A}_i$. These UEs have temporarily been allocated $|\mathcal{A}_i|$ mutually orthogonal pilot sequences, which however are reused in other cells (using some reuse factor that guarantees low pilot contamination [5]). The payload data blocks can be operated as in the classical Massive MIMO works [4]–[6], [8]–[11], which provide methods to achieve high data rates.

The second category is dedicated for RA from inactive UEs (i.e., some of those in $\mathcal{U}_i \setminus \mathcal{A}_i$) that wish to be admitted to the payload data blocks; that is, to be allocated a temporary dedicated pilot. This category has not been studied in the Massive MIMO context and is the main focus of this paper. The time-frequency location of the RA blocks is different between adjacent cells, as illustrated in Fig. 2. This design choice protects the RA procedure from the strongest forms of inter-cell interference, as will be shown later.

A. Strongest-User Collision Resolution (SUCRe)—Overview

The key contribution of this paper is the strongest-user collision resolution (SUCRe) protocol, which is an efficient way to operate the RA blocks in beyond-LTE Massive MIMO systems. We first give a brief overview of the protocol and then provide the exact analytical details. The four main steps of the SUCRe protocol are illustrated in Fig. 3. There is also a preliminary Step 0 in which the BS broadcasts a control signal. Each UE uses this signal to estimate its average channel gain and to synchronize itself towards the BS. In OFDM, the UE and the BS need to be synchronized in frequency and the timing delay can be neglected if it is shorter than the cyclic prefix. The round-trip time determines the maximum timing delay, thus the normal CP in LTE allows for 750 m cell radius and the extended CP allows for 2.5 km—these are substantially larger than the 250–500 m cell radius typical in urban deployments. This paper focuses on such urban scenarios and we stress that the spatial multiplexing in Massive MIMO is ideal for crowded urban settings. In Step 1, a subset
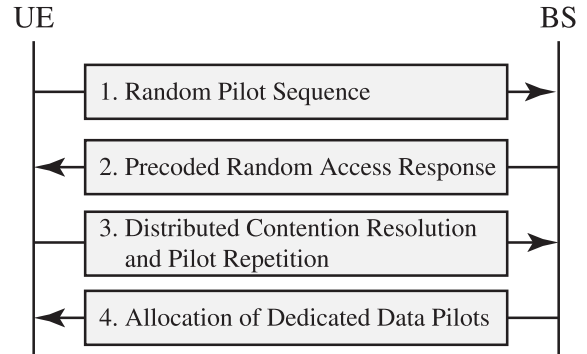

Fig. 3. Proposed SUCRe random access protocol for Massive MIMO.

of the inactive UEs in cell i wants to become active. Each such UE selects a pilot sequence at random from a predefined pool of RA pilots. BS i estimates the channel that each pilot has propagated over. If multiple UEs selected the same RA pilot, a collision has occurred and the BS obtains an estimate of the superposition of the UE channels. The BS cannot detect if collisions occurred at this point, which resembles the situation in LTE.

In Step 2, the BS responds by sending DL pilots that are precoded using the channel estimates, which results in spatially directed signals toward the UEs that sent the particular RA pilot. The DL signal features an array gain of M that is divided between the UEs that sent the RA pilot. Due to channel reciprocity, the share of the array gain is proportional to their respective UL signal gains, particularly when M is large, which enables each UE to estimate the sum of the signal gains and compare it with its own signal gain (using information obtained in Step 0). Each UE can thereby detect RA collisions in a *distributed way*. This departs from the conventional approach in which collisions are detected in a *centralized way* at the BS and broadcasted to the UEs.

Based on the detection in Step 2 and the favorable propagation of Massive MIMO channels, the UEs can resolve RA contentions in Step 3 by applying a local decision rule: only the UE with the strongest UL signal gain should repeat the RA pilot. This is a key advantage over LTE, where all contending UEs repeat the preamble in Step 3. This mechanism is illustrated in Fig. 4. The probability of non-colliding pilot transmission in Step 3 is vastly increased in the
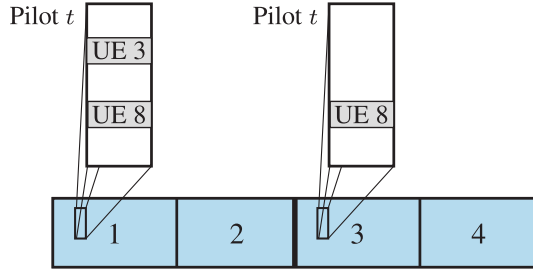
Fig. 4. Illustration of a UE collision at RA pilot t resolved by the SUCRe protocol, where only the UE with strongest signal gain repeats the pilot in Step 3.

SUCRe protocol, which enables the network to admit UEs also in crowded scenarios. The transmission in Step 3 also contains the identity of the UE and a request for payload transmission, resembling the RRC Connection Request in LTE. Step 4 grants these resources by assigning a pilot sequence that can be used in the payload blocks or starts further contention resolution (e.g., in an LTE fashion or by repeating the SUCRe protocol) in the few cases when RA collisions remain. Hence, the SUCRe protocol both stands on its own and can complement conventional contention resolution methods.

B. Detailed Description of the SUCRe Protocol

Next, we describe and analyze the proposed RA protocol in detail. Without loss of generality, we focus on an arbitrary cell, say cell 0, and consider how interference from other cells impacts the operation. Let $\mathcal{K}_0 = \mathcal{U}_0 \setminus \mathcal{A}_0$ denote the set of inactive UEs in cell 0 in a given RA block. Hence, there are $K_0 = |\mathcal{K}_0|$ inactive UEs in cell 0. These are assumed to share τ_p mutually orthogonal RA pilot sequences $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{\tau_p} \in \mathbb{C}^{\tau_p}$ that span τ_p UL channel uses and satisfy $\|\boldsymbol{\psi}_t\|^2 = \tau_p$. The inactive UEs are not fully time-synchronized, but the pilot orthogonality is maintained at the receiver since we consider urban scenarios where the roundtrip delays are smaller than the cyclic prefix. We typically have $\tau_p \ll K_0$, but there is no formal constraint.

Each of the K_0 UEs picks one of the τ_p pilots uniformly at random in each RA block: UE k selects pilot $c(k) \in \{1, 2, \ldots, \tau_p\}$. Furthermore, each UE would like to become active in the current block with probability $P_a \leq 1$, which is a fixed scenario-dependent parameter that describes how often a UE has data packets to transmit or receive. An access attempt by UE k consists of transmitting the pilot $\boldsymbol{\psi}_{c(k)}$ with a non-zero power $\rho_k > 0$, otherwise it stays silent by setting $\rho_k = 0$. Hence, each inactive UE will transmit a particular pilot sequence $\boldsymbol{\psi}_t$ (using non-zero power) with probability P_a/τ_p. The set $\mathcal{S}_t = \{k : c(k) = t, \rho_k > 0\}$ contains the indices of the UEs that transmit pilot t. Based on this model, the number of UEs, $|\mathcal{S}_t|$, that transmits $\boldsymbol{\psi}_t$ has a binomial distribution[1]:

$$|\mathcal{S}_t| \sim B\left(K_0, \frac{P_a}{\tau_p}\right). \tag{1}$$

We notice that pilot t is unused ($|\mathcal{S}_t| = 0$) with probability $(1 - \frac{P_a}{\tau_p})^{K_0}$ and selected by only one UE ($|\mathcal{S}_t| = 1$) with probability $K_0 \frac{P_a}{\tau_p}(1 - \frac{P_a}{\tau_p})^{K_0-1}$. Consequently, an RA collision ($|\mathcal{S}_t| \geq 2$) occurs at this arbitrary pilot with probability

$$1 - \left(1 - \frac{P_a}{\tau_p}\right)^{K_0} - K_0 \frac{P_a}{\tau_p}\left(1 - \frac{P_a}{\tau_p}\right)^{K_0-1}. \tag{2}$$

These collisions need to be detected and resolved before any UE can be admitted into the payload blocks. The SUCRe protocol is a distributed method to resolve pilot collisions at the UE side by utilizing properties of Massive MIMO channels.

The channel vector between UE $k \in \mathcal{K}_0$ and its BS is denoted by $\mathbf{h}_k \in \mathbb{C}^M$. We adopt a very general propagation model where the channels are assumed to satisfy the following two conditions (almost surely):

$$\frac{\|\mathbf{h}_k\|^2}{M} \xrightarrow{M \to \infty} \beta_k, \quad \forall k, \tag{3}$$

$$\frac{\mathbf{h}_k^{\mathrm{H}} \mathbf{h}_i}{M} \xrightarrow{M \to \infty} 0, \quad \forall k, i, \ k \neq i, \tag{4}$$

for some strictly positive value of β_k that is known to UE k (it was estimated in Step 0). Such channels are said to offer *channel hardening* and *asymptotic favorable propagation* [25].

The properties (3) and (4) are satisfied (almost surely) by a variety of stochastic channel models; for example, when $\mathbf{h}_k = \mathbf{R}_k^{1/2}\mathbf{x}_k$ where $\mathbf{R}_k \in \mathbb{C}^{M \times M}$ is a positive semi-definite matrix with bounded spectral norm and $\mathbf{x}_k \in \mathbb{C}^M$ has i.i.d. entries with zero mean and bounded eighth-order moment [26, Ths. 3.4 and 3.7]. In this case we have $\mathrm{tr}(\mathbf{R}_k)/M \to \beta_k$. Asymptotic favorable propagation can also be obtained for deterministic line-of-sight channels; for example, for uniform linear arrays (ULAs) where the UEs have distinct angles with respect to the BS array [25].

We will describe the four steps of the SUCRe protocol under the assumption that the channels satisfy (3) and (4). In Section III, we particularize the protocol for uncorrelated Rayleigh fading channels, which enables us to quantify the performance in further detail.

1) Random Pilot Sequence: In Step 1 of the proposed RA protocol, the BS receives the signal $\mathbf{Y} \in \mathbb{C}^{M \times \tau_p}$ from the pilot transmission:

$$\mathbf{Y} = \sum_{k \in \mathcal{K}_0} \sqrt{\rho_k}\mathbf{h}_k \boldsymbol{\psi}_{c(k)}^{\mathrm{T}} + \mathbf{W} + \mathbf{N}, \tag{5}$$

where $\mathbf{N} \in \mathbb{C}^{M \times \tau_p}$ is independent receiver noise with each element distributed as $\mathcal{CN}(0, \sigma^2)$. The matrix $\mathbf{W} \in \mathbb{C}^{M \times \tau_p}$ represents interference from other cells.

By correlating $\mathbf{Y}$ with an arbitrary (normalized) pilot sequence $\boldsymbol{\psi}_t$, the BS obtains

$$\mathbf{y}_t = \mathbf{Y}\frac{\boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|} = \sum_{i \in \mathcal{S}_t} \sqrt{\rho_i}\|\boldsymbol{\psi}_t\|\mathbf{h}_i + \mathbf{W}\frac{\boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|} + \mathbf{n}_t$$

$$= \sum_{i \in \mathcal{S}_t} \sqrt{\rho_i \tau_p}\mathbf{h}_i + \mathbf{W}\frac{\boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|} + \mathbf{n}_t, \tag{6}$$

---

[1]This is a simplified model where each RA block is treated independently. However, a UE that fails to access the network in one block will soon try again, which creates a correlation between RA blocks. This practical scenario is studied numerically in Section IV and we stress that the proposed protocol can be applied under any UE distribution.

where $\mathbf{n}_t = \mathbf{N}\frac{\boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_M)$ is the effective receiver noise and we notice that $\|\boldsymbol{\psi}_t\| = \sqrt{\tau_p}$. Recall that $\mathcal{S}_t$ is the set of UEs that transmitted pilot $\boldsymbol{\psi}_t$.

The inter-cell interference $\mathbf{W}$ can be modeled as

$$\mathbf{W} = \sum_l \mathbf{w}_l \mathbf{d}_l^{\mathrm{T}} + \sum_{t=1}^{\tau_p} \sum_{k \in \mathcal{S}_t^{\mathrm{interf}}} \sqrt{\rho_{t,k}} \mathbf{g}_{t,k} \boldsymbol{\psi}_t^{\mathrm{T}}. \tag{7}$$

The first summation in (7) is over the interfering data transmissions carried out in neighboring cells with another color than cell 0 in Fig. 2. The lth interferer has the channel $\mathbf{w}_l \in \mathbb{C}^M$ to the BS in cell 0 and transmits some random data sequence $\mathbf{d}_l \in \mathbb{C}^{\tau_p}$. The second summation is over the interferers in cells with the same color as cell 0 in Fig. 2, which also perform RA. The interferers that use pilot $\boldsymbol{\psi}_t$ are gathered in the set $\mathcal{S}_t^{\mathrm{interf}}$, and member $k \in \mathcal{S}_t^{\mathrm{interf}}$ has the channel $\mathbf{g}_{t,k}$ to the BS in cell 0 and uses the transmit power $\rho_{t,k}$. It follows that

$$\mathbf{W}\frac{\boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|} = \sum_l \mathbf{w}_l \frac{\mathbf{d}_l^{\mathrm{T}} \boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|} + \sum_{k \in \mathcal{S}_t^{\mathrm{interf}}} \sqrt{\rho_{t,k}\tau_p}\mathbf{g}_{t,k}. \tag{8}$$

Assuming that all the interfering channels also satisfy the conditions in (3) and (4), denoted as $\|\mathbf{w}_l\|^2/M \to \beta_{w,l}$ and $\|\mathbf{g}_{t,k}\|^2/M \to \beta_{t,k}$, we obtain

$$\frac{\left|\mathbf{W}\frac{\boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|}\right|^2}{M} \to \underbrace{\sum_l \beta_{w,l}\frac{|\mathbf{d}_l^{\mathrm{T}}\boldsymbol{\psi}_t^*|^2}{\|\boldsymbol{\psi}_t\|^2} + \sum_{k \in \mathcal{S}_t^{\mathrm{interf}}} \rho_{t,k}\tau_p\beta_{t,k}}_{\omega_t} \tag{9}$$

as $M \to \infty$. Note that there is interference in ω_t from both data transmission and RA pilots in other cells, but the former typically dominates since the closest neighboring cells in Fig. 2 transmits data. It is only the value of ω_t that is important in the remainder of this section, and not how it is computed.

Remark 1 (Detecting Active Pilots): The BS can utilize $\mathbf{y}_t \in \mathbb{C}^M$ to determine if $|\mathcal{S}_t| \geq 1$ or $|\mathcal{S}_t| = 0$ for the considered pilot sequence; that is, whether or not there is at least one active UE. This is particularly easy in Massive MIMO systems that operate under channel hardening and asymptotic favorable propagation since

$$\frac{\|\mathbf{y}_t\|^2}{M} \xrightarrow{M \to \infty} \sum_{i \in \mathcal{S}_t} \rho_i\beta_i\tau_p + \omega_t + \sigma^2. \tag{10}$$

This is an additional feature since the SUCRe protocol does not require the BS to know which pilots were used in Step 1.

2) Precoded Random Access Response: In Step 2, the BS responds to the RA pilots by sending orthogonal precoded DL pilot signals that correspond to each of the RA pilots that were used in the UL. The response to $\boldsymbol{\psi}_t$ is a pilot sequence $\boldsymbol{\phi}_t \in \mathbb{C}^{\tau_p}$, and the DL pilot sequences $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{\tau_p} \in \mathbb{C}^{\tau_p}$ are mutually orthogonal and satisfy $\|\boldsymbol{\phi}_t\|^2 = \tau_p$. Note that $\mathbf{y}_t$ in (6) is a weighted superposition of the channels of UEs that used pilot t (it is also impaired by interference and noise). If the BS uses the normalized conjugate $\mathbf{y}_t^*/\|\mathbf{y}_t\|$ as precoding vector when sending the pilot sequence $\boldsymbol{\phi}_t$ in the DL, the signal will be directed in a multi-cast maximum ratio transmission

fashion towards the UEs in $\mathcal{S}_t$. The complete precoded DL pilot signal $\mathbf{V} \in \mathbb{C}^{M \times \tau_p}$ is

$$\mathbf{V} = \sqrt{q} \sum_{t=1}^{\tau_p} \frac{\mathbf{y}_t^*}{\|\mathbf{y}_t\|}\boldsymbol{\phi}_t^{\mathrm{T}}, \tag{11}$$

where the DL transmit power q has a predefined value. Note that all pilot sequences are also sent in the DL and the pilot length τ_p is independent of the number of antennas.

The received signal $\mathbf{z}_k \in \mathbb{C}^{\tau_p}$ at UE $k \in \mathcal{S}_t$ is

$$\mathbf{z}_k^{\mathrm{T}} = \mathbf{h}_k^{\mathrm{T}}\mathbf{V} + \boldsymbol{v}_k^{\mathrm{T}} + \boldsymbol{\eta}_k^{\mathrm{T}}, \tag{12}$$

where $\mathbf{h}_k^{\mathrm{T}}$ is the reciprocal DL channel, $\boldsymbol{v}_k \in \mathbb{C}^{\tau_p}$ is inter-cell interference, and $\boldsymbol{\eta}_k \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}_{\tau_p})$ is receiver noise. By correlating the received signal $\mathbf{z}_k$ with the (normalized) DL pilot sequence $\boldsymbol{\phi}_t$, the UE obtains

$$z_k = \mathbf{z}_k^{\mathrm{T}}\frac{\boldsymbol{\phi}_t^*}{\|\boldsymbol{\phi}_t\|} = \sqrt{q\tau_p}\mathbf{h}_k^{\mathrm{T}}\frac{\mathbf{y}_t^*}{\|\mathbf{y}_t\|} + \boldsymbol{v}_k^{\mathrm{T}}\frac{\boldsymbol{\phi}_t^*}{\|\boldsymbol{\phi}_t\|} + \eta_k \tag{13}$$

where $\eta_k = \boldsymbol{\eta}_k^{\mathrm{T}}\frac{\boldsymbol{\phi}_t^*}{\|\boldsymbol{\phi}_t\|} \sim \mathcal{CN}(0, \sigma^2)$ is the effective receiver noise. We notice that

$$\frac{z_k}{\sqrt{M}} = \sqrt{q\tau_p}\frac{(\mathbf{h}_k^{\mathrm{H}}\mathbf{y}_t)^*}{M}\frac{1}{\sqrt{\frac{1}{M}\|\mathbf{y}_t\|^2}} + \frac{\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{\phi}_t^*}{\sqrt{M}\|\boldsymbol{\phi}_t\|} + \frac{\eta_k}{\sqrt{M}}$$

$$\xrightarrow{M \to \infty} \frac{\sqrt{\rho_k q}\beta_k\tau_p}{\sqrt{\sum_{i \in \mathcal{S}_t} \rho_i\beta_i\tau_p + \omega_t + \sigma^2}} \tag{14}$$

by utilizing the asymptotic favorable propagation, the convergence in (10), exploiting the fact that the noise does not increase with M, and assuming that the inter-cell interference $\boldsymbol{v}_k$ is unaffected by M. The latter assumption is well motivated if the closest interfering cells that send RA pilots assign the DL pilots in different ways, to avoid causing coherent pilot contaminated interference.[2]

Looking at (14), let us define

$$\alpha_t = \sum_{i \in \mathcal{S}_t} \rho_i\beta_i\tau_p + \omega_t \tag{15}$$

as the sum of the signal and interference gains received at the BS during the UL transmission of pilot $\boldsymbol{\psi}_t$. The intra-cell signals are amplified by a factor τ_p, as compared to the inter-cell interference, which is the processing gain from having a pilot sequence that spans τ_p symbols. Let the function $\Re(\cdot)$ give the real part of its input. Based on (14), we obtain the approximation

$$\frac{\Re(z_k)}{\sqrt{M}} \approx \frac{\sqrt{\rho_k q}\beta_k\tau_p}{\sqrt{\alpha_t + \sigma^2}}, \tag{16}$$

where we discard the imaginary part of z_k that only contains noise, interference and estimation errors. UE k can use this approximation to estimate α_t:

$$\hat{\alpha}_{t,k}^{\mathrm{approx1}} = \max\left(\frac{Mq\rho_k\beta_k^2\tau_p^2}{(\Re(z_k))^2} - \sigma^2, \rho_k\beta_k\tau_p\right), \tag{17}$$

---

[2]For example, one cell might assign $\boldsymbol{\phi}_t = \boldsymbol{\psi}_t$, the next one assigns $\boldsymbol{\phi}_t = \boldsymbol{\psi}_{t+1}$, and another one assigns $\boldsymbol{\phi}_t = \boldsymbol{\psi}_{t+2}$. With such pilot switching, the closest τ_p cells that perform RA will not have any UEs that collide in both the UL and DL, which alleviates the coherent pilot contaminated interference between these cells.

where $\max(\cdot, \cdot)$ takes the maximum of the two values (since UE k knows that $\alpha_t \geq \rho_k \beta_k \tau_p$). This estimator is asymptotically error-free, as $M \to \infty$, due to (14).

3) Contention Resolution & Pilot Repetition: The UL pilot transmission is repeated in Step 3. The main goal of the proposed protocol is to resolve pilot contentions in a distributed manner in Step 3, so that each pilot is only repeated by one UE. Each UE $k \in \mathcal{S}_t$ knows its own average signal gain $\rho_k \beta_k \tau_p$ and has an estimate $\hat{\alpha}_{t,k}$ of the sum of the signal gains of the contending UEs (plus inter-cell interference), such that it can infer:

- If a pilot collision has occurred: $\hat{\alpha}_{t,k} > \rho_k \beta_k \tau_p$ (with a margin that accounts for inter-cell interference);
- How strong its own signal is relative to the sum of all the contenders' signals: $\rho_k \beta_k \tau_p / \hat{\alpha}_{t,k}$.

Since the number of contenders, $|\mathcal{S}_t|$, is unknown, a UE can only compare its own signal gain with the summation of the gains of its contenders. To resolve the contention we make the following definition.

Definition 1: The contention winner is the UE $k \in \mathcal{S}_t$ with the largest $\rho_k \beta_k \tau_p$, referred to as the *strongest user*.

In the asymptotic case when α_t is exactly known, UE k is sure to be the contention winner if $\rho_k \beta_k \tau_p > \alpha_t - \rho_k \beta_k \tau_p$, irrespective of how many contenders there are. This criterion can be written as $\rho_k \beta_k \tau_p > \alpha_t/2$ and interpreted as having a UE k with a signal gain that is greater than the sum of all the other UEs' signal gains.

Definition 2: We have *resolved a collision* if and only if a single UE appoints itself the contention winner.

An example where a two-UE collision is resolved is given in Fig. 4. If there are $|\mathcal{S}_t| = 2$ UEs (or $|\mathcal{S}_t| = 1$ for that matter) the criterion $\rho_k \beta_k \tau_p > \alpha_t/2$ can be used to resolve any such collision, in the special case when α_t is known and there is no inter-cell interference. In practice, only the estimate $\hat{\alpha}_{t,k}$ is known, it happens that $|\mathcal{S}_t| \geq 3$ and there will be inter-cell interference. There is then a risk that multiple UEs or no UE identify themselves as the contention winner.

Definition 3: A *false negative* occurs when none of the colliding UEs identifies itself as the contention winner. A *false positive* occurs when more than one colliding UE identify itself as the contention winner.

We propose that each UE $k \in \mathcal{S}_t$ applies the following distributed *decision rule*:

$$\mathcal{R}_k: \quad \rho_k \beta_k \tau_p > \hat{\alpha}_{t,k}/2 + \epsilon_k \quad \text{(repeat)}, \tag{18}$$

$$I_k: \quad \rho_k \beta_k \tau_p \leq \hat{\alpha}_{t,k}/2 + \epsilon_k \quad \text{(inactive)}. \tag{19}$$

UE $k \in \mathcal{S}_t$ concludes that it has the strongest signal gain if $\mathcal{R}_k$ is true and *repeats* the transmission of pilot $\boldsymbol{\psi}_t$ in Step 3. If it instead concludes that I_k is true, it decides to remain *inactive* by pulling out from the RA attempt and try again later. The estimation errors and inter-cell interference can cause false positives or negatives. The bias parameter $\epsilon_k \in \mathbb{R}$ can be used to tune the system behavior to the final performance criterion; for example, to maximize the average number of resolved collisions or to minimize the risk of false positives (or negatives).

The probability of resolving a contention is determined by the decision rule. By numbering the active UEs in $\mathcal{S}_t$ from 1 to $|\mathcal{S}_t|$, the probability of resolving a pilot collision with $|\mathcal{S}_t|$ contenders is

$$\begin{aligned} P_{|\mathcal{S}_t|,\text{resolved}} = {} & \Pr\{\mathcal{R}_1, I_2, \ldots, I_{|\mathcal{S}_t|}\} \\ & + \Pr\{I_1, \mathcal{R}_2, I_3, \ldots, I_{|\mathcal{S}_t|}\} + \cdots \\ & + \Pr\{I_1, \ldots, I_{|\mathcal{S}_t|-1}, \mathcal{R}_{|\mathcal{S}_t|}\}, \end{aligned} \tag{20}$$

where the randomness is due to channel realizations, inter-cell interference, and noise (and possibly also random UE locations). In the special case $|\mathcal{S}_t| = 2$, (20) reduces to

$$P_{2,\text{resolved}} = \Pr\{\mathcal{R}_1, I_2\} + \Pr\{I_1, \mathcal{R}_2\}, \tag{21}$$

while a false negative occurs if both the UEs pull out (with probability $\Pr\{I_1, I_2\}$) and a false positive occurs when both UEs repeat the pilot (with probability $\Pr\{\mathcal{R}_1, \mathcal{R}_2\}$).

Remark 2 (Probabilistic Bias Terms): The decision rule in (18) appears to make a hard decision on whether or not UE k is the strongest user. This can, however, be softened by using random bias terms. For example, UE k might know that a certain ratio $\rho_k \beta_k \tau_p / \hat{\alpha}_{t,k}$ implies a certain probability of being the strongest UE. The bias term can then be made a random variable that depends on this ratio and makes the UE appoint itself the strongest user with the same probability (or a modified probability that, for example, prioritizes weaker UEs over stronger ones). Since the exact details depend strongly on the propagation environment and UE distribution, which are hard to compute and model exactly for practical setups, the design of probabilistic bias terms is mainly an engineering problem that is not studied further in this paper.

4) Allocation of Dedicated Payload Pilots: The BS receives the repeated RA pilot transmissions in Step 3, which are followed by UL messages that, for example, can contain the unique identity number of the UE. The BS uses the tth pilot signal to estimate the channel to the UE (or UEs) that sent $\boldsymbol{\psi}_t$ in Step 3 and tries to decode the corresponding message. If the decoding is successful, the BS has identified one of the UEs in $\mathcal{S}_t$ and can admit it to the payload coherence blocks by allocating a pilot sequence (which typically is unique within the cell). This resource allocation decision is transmitted in the DL in Step 4, similarly to the precoded response in Step 2. The transmission can also contain other important information for the subsequent data transmission, such as timing advance. If the decoding fails, the SUCRe protocol has failed to resolve the collision. Note that if pilot t was unused in Step 1 (i.e., $|\mathcal{S}_t| = 0$), it will also be unused in Step 3 and hence the "decoding" will fail—there is no need for the BS to explicitly determine that $|\mathcal{S}_t| = 0$.

The SUCRe protocol is repeated at a given interval. The UEs that were not admitted in Step 4 can be instructed when and how to transmit new RA pilots, for example, after a random waiting time. Alternatively, we can add additional steps on top of the SUCRe protocol to resolve the remaining collisions, by utilizing any conventional contention resolution method. For example, the UEs that collided in Step 3 can select new UL pilots at random and the risk of new collisions is vastly reduced since there are few remaining collisions.

Remark 3 (Fairness): If the UEs transmit at constant power, then the SUCRe protocol gives priority to UEs with

the strong channel gains, which are typically in the cell center, while cell-edge UEs have weaker channel gains and are more likely to lose a contention. This short-term fairness issue is the price to pay for being able to resolve many collisions; the simulations in Section IV show that around $0.9\tau_p$ UEs out of the maximum τ_p are admitted in each RA block—even under high load. Since only a few UEs need to make new attempts, there will be fewer collisions in the long-run and also the weakest UEs will succeed after a few attempts (see Section IV-D for numerical evidence). If more short-term fairness is desired (e.g., for more rapid handover), the decision rule can be changed towards this end; for example, by using probabilistic bias terms (see Remark 2) and/or uplink power control (see Remark 4 below).

III. PERFORMANCE WITH UNCORRELATED RAYLEIGH FADING CHANNELS

In this section, we consider the special case of uncorrelated Rayleigh fading channels with

$$\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{I}_M) \tag{22}$$

for all UEs $k \in \mathcal{K}_0$. Furthermore, the inter-cell interference term is modeled as $\boldsymbol{v}_k \sim \mathcal{CN}(\mathbf{0}, \Upsilon_k \mathbf{I}_{\tau_p})$ and is independent of the other signals. This model allows for a relatively tractable performance analysis for any value of M, in contrast to the analysis in Section II that focused on M for which the channel hardening and asymptotic favorable propagation properties are applicable (typically: $M > 50$). Recall that the received signal $z_k \in \mathbb{C}$ at UE $k \in \mathcal{S}_t$ in Step 2 was given in (13). The following lemma characterizes the distribution of this random variable.

Lemma 1: Consider uncorrelated Rayleigh fading channels. For any UE $k \in \mathcal{S}_t$ the received signal in (13) can be expressed as $z_k = g_k + v_k$, where

$$g_k = \sqrt{\frac{1}{2}\frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}} x, \quad x \sim \chi_{2M} \tag{23}$$

$$v_k \sim \mathcal{CN}\left(0, \sigma^2 + \Upsilon_k + q\beta_k\tau_p - \frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}\right) \tag{24}$$

are independent and χ_n denotes a chi-distribution with n degrees of freedom.

Proof: The proof is given in Appendix B. ∎

By using the statistical properties in Lemma 1, we can compute the mean and variance of the normalized received DL signal $z_k/\sqrt{M}$:

$$\mathbb{E}\left\{\frac{z_k}{\sqrt{M}}\right\} = \sqrt{\frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}} \frac{\Gamma\left(M+\frac{1}{2}\right)}{\sqrt{M}\Gamma(M)}, \tag{25}$$

$$\mathbb{V}\left\{\frac{z_k}{\sqrt{M}}\right\} = \frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}\left(1 - \left(\frac{\Gamma\left(M+\frac{1}{2}\right)}{\sqrt{M}\Gamma(M)}\right)^2\right)$$
$$+ \frac{1}{M}\left(\sigma^2 + \Upsilon_k + q\beta_k\tau_p - \frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}\right). \tag{26}$$

By taking the limit $M \to \infty$ and treating the Gamma function using Lemma 2 in Appendix A, we obtain

$$\mathbb{E}\left\{\frac{z_k}{\sqrt{M}}\right\} \to \sqrt{\frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}}, \quad \text{as} \quad M \to \infty, \tag{27}$$

$$\mathbb{V}\left\{\frac{z_k}{\sqrt{M}}\right\} \to 0, \quad \text{as} \quad M \to \infty. \tag{28}$$

The mean value approaches the limit in (14) and the variance goes to zero, which confirms that Rayleigh fading channels offer asymptotic favorable propagation.

A. Different Estimators of α_t

Instead of exploiting the channel hardening and asymptotic favorable propagation to estimate α_t, as we did in (17), we can use the exact statistics from Lemma 1 to obtain the maximum likelihood (ML) estimate.

Theorem 1: Consider uncorrelated Rayleigh fading channels. The ML estimate of α_t from the observation $z_k = z_{k,\Re} + J z_{k,\Im}$ (with $z_{k,\Re}, z_{k,\Im} \in \mathbb{R}$) is

$$\hat{\alpha}_{t,k}^{\text{ML}} = \underset{\alpha \geq \rho_k \beta_k \tau_p}{\arg\max}\, f_1\left(z_{k,\Re}|\alpha\right) f_2\left(z_{k,\Im}|\alpha\right) \tag{29}$$

for the conditional probability density functions (PDFs)

$$f_1\left(z_{k,\Re}|\alpha\right) = \frac{e^{-\frac{(z_{k,\Re})^2}{\lambda_2}\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)}}{\Gamma(M)\lambda_1^M \sqrt{\pi \lambda_2}} \sum_{n=0}^{2M-1}\binom{2M-1}{n}$$
$$\times \frac{\left(\Gamma\left(\frac{n+1}{2}\right) + c_n(z_{k,\Re})\gamma\left(\frac{n+1}{2}, \frac{(z_{k,\Re})^2}{\lambda_2}\frac{\lambda_1}{\lambda_1+\lambda_2}\right)\right)}{\left(\frac{z_{k,\Re}}{\lambda_2}\right)^{n+1-2M}\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right)^{2M-\frac{n+1}{2}}} \tag{30}$$

$$f_2\left(z_{k,\Im}|\alpha\right) = \frac{1}{\sqrt{\pi \lambda_2}}e^{-\frac{(z_{k,\Im})^2}{\lambda_2}}, \tag{31}$$

where $\gamma(\cdot, \cdot)$ is the lower incomplete gamma function,[3]

$$c_n(z) = \begin{cases} (-1)^n & z \geq 0, \\ -1 & z < 0, \end{cases} \tag{32}$$

and the coefficients λ_1 and λ_2 depend on α as

$$\lambda_1 = \frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha + \sigma^2} \tag{33}$$

$$\lambda_2 = \sigma^2 + \Upsilon_k + q\beta_k\tau_p - \lambda_1. \tag{34}$$

Proof: The proof is given in Appendix B. ∎

The ML estimate $\hat{\alpha}_{t,k}^{\text{ML}}$ of α_t can be computed numerically using Theorem 1.[4] An approximate estimate in closed-form can also be obtained from (25) by utilizing the fact that

$$\Re(z_k) \approx \mathbb{E}\{z_k\} = \sqrt{\frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}}\frac{\Gamma\left(M+\frac{1}{2}\right)}{\Gamma(M)} \tag{35}$$

---

[3]For any positive integer m, the lower incomplete gamma function can be computed as $\gamma(m, x) = \Gamma(m) - \Gamma(m)e^{-x}\sum_{k=0}^{m-1} x^k / \Gamma(k+1)$.

[4]The conditional PDF $f_1\left(z_{k,\Re}|\alpha\right)$ contains several terms that grow rapidly with M, while their ratios remain small. Hence, a careful implementation of the PDF is needed for numerical stability. The simulations in this paper were implemented successfully by taking the logarithm of each term in the summation.
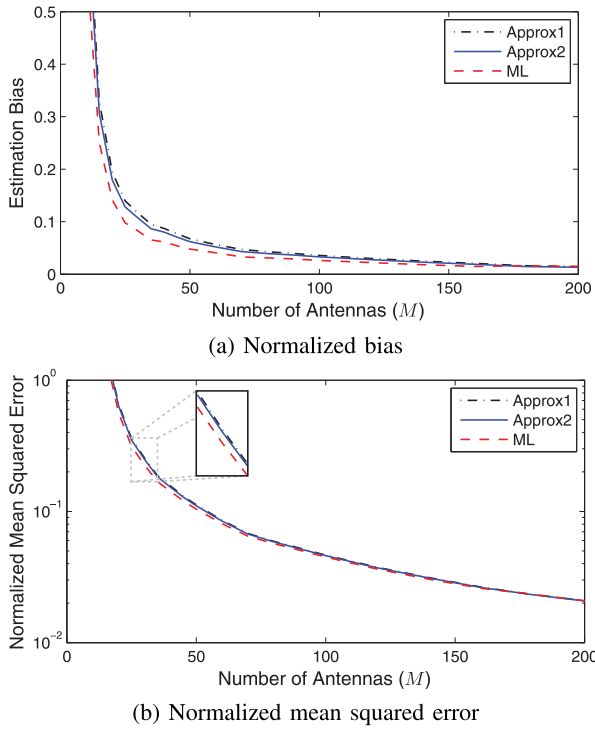
(a) Normalized bias

(b) Normalized mean squared error

Fig. 5. Comparison of three estimators of the signal gain α_t: $\hat{\alpha}_{t,k}^{\mathrm{ML}}$, $\hat{\alpha}_{t,k}^{\mathrm{approx1}}$, and $\hat{\alpha}_{t,k}^{\mathrm{approx2}}$. The true value is $\alpha_t = 20$, while the pilot SNR at UE i is 10 dB.

when M is large. By solving the equation $\Re(z_k) = \mathbb{E}\{z_k\}$ for α_t we obtain the estimator

$$\hat{\alpha}_{t,k}^{\mathrm{approx2}} = \max\left(\left(\frac{\Gamma\left(M+\frac{1}{2}\right)}{\Gamma(M)}\right)^2 \frac{q\rho_k\beta_k^2\tau_p^2}{(\Re(z_k))^2} - \sigma^2, \, \rho_k\beta_k\tau_p\right) \tag{36}$$

since UE k knows that $\alpha_t \geq \rho_k\beta_k\tau_p$. This estimator is slightly different from the one obtained in (17), but they are asymptotically equivalent as $M \to \infty$ due to (45) and both are asymptotically error free.

We have obtained three different estimators of α_t: $\hat{\alpha}_{t,k}^{\mathrm{ML}}$, $\hat{\alpha}_{t,k}^{\mathrm{approx1}}$, and $\hat{\alpha}_{t,k}^{\mathrm{approx2}}$. Which one is preferred in practice? The performance of these estimators is compared in Fig. 5, where $\alpha_t = 20$ and UE i estimates α_t while having $q = \rho_i\beta_i = \sigma^2 = 1$ and $\tau_p = 10$. This corresponds to an SNR of 0 dB, while the effective pilot SNR is 10 dB. Fig. 5(a) shows the normalized bias $(\mathbb{E}\{\hat{\alpha}_{t,k}\} - \alpha_t)/\alpha_t$ and Fig. 5(b) shows the normalized mean-squared error (NMSE) $\mathbb{E}\{|\hat{\alpha}_{t,k} - \alpha_t|^2\}/\alpha_t$. All three estimators perform badly for $M < 25$, but become asymptotically unbiased as M increases and achieve NMSEs below 10^{-1} for $M \geq 50$. The SUCRe protocol is thus particularly useful in Massive MIMO. All estimators have a tendency to overestimate α_t, as seen from the positive bias. The ML estimator $\hat{\alpha}_{t,k}^{\mathrm{ML}}$ provides the smallest NMSEs, as expected, while $\hat{\alpha}_{t,k}^{\mathrm{approx2}}$ is better than $\hat{\alpha}_{t,k}^{\mathrm{approx1}}$. However, the differences are tiny and thus we appoint $\hat{\alpha}_{t,k}^{\mathrm{approx2}}$ as the preferred choice since the ML estimator is computationally involved.

B. Probability of Pilot Repetition

The core of the SUCRe protocol is that only one UE should transmit pilot t in Step 3. The probability that a particular UE decides to repeat its pilot transmission is computed as follows.

Theorem 2: If UE $k \in \mathcal{S}_t$ estimates α_t as in (36) and applies the proposed decision rule with[5] $\epsilon_k < \rho_k\beta_k\tau_p/2$, then the probability of repeating the pilot in Step 3 is

$$\Pr\{\mathcal{R}_k\} = 1 - \Pr\{\Re(z_k) \leq \sqrt{\zeta_k}\} + \Pr\{\Re(z_k) \leq -\sqrt{\zeta_k}\} \tag{37}$$

where

$$\zeta_k = \left(\frac{\Gamma\left(M+\frac{1}{2}\right)}{\Gamma(M)}\right)^2 \frac{q\rho_k\beta_k^2\tau_p^2}{\sigma^2 + 2(\rho_k\beta_k\tau_p - \epsilon_k)} \tag{38}$$

and

$$\Pr\{\Re(z_k) \leq b\}$$
$$= Q\left(-\frac{b\sqrt{2}}{\sqrt{\lambda_2}}\right) - \sum_{k=0}^{M-1} \frac{e^{-\frac{b^2}{\lambda_2}\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)}}{\Gamma(k+1)\lambda_1^k\sqrt{\pi\lambda_2}}$$
$$\times \sum_{n=0}^{2k} \binom{2k}{n} \frac{\left(\Gamma\left(\frac{n+1}{2}\right) + c_n(b)\gamma\left(\frac{n+1}{2}, \frac{b^2}{\lambda_2}\frac{\lambda_1}{\lambda_1+\lambda_2}\right)\right)}{2\left(\frac{b}{\lambda_2}\right)^{n-2k}\left(\frac{1}{\lambda_1}+\frac{1}{\lambda_2}\right)^{2k+\frac{1}{2}-\frac{n}{2}}}. \tag{39}$$

In these expressions, $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-t^2/2}dt$ is the Q-function, $c_n(\cdot)$ is defined in (32), and λ_1 and λ_2 are given in (33)–(34) with $\alpha = \alpha_t$.

Proof: The proof is given in Appendix B. ∎

This theorem provides the CDF of $\Re(z_k)$ and uses it to compute the exact probability $\Pr\{\mathcal{R}_k\}$ that UE k repeats the pilot in Step 3. Note that $\Pr\{\Re(z_k) \leq -\sqrt{\zeta_k}\} > 0$ in general, since there is a (tiny) probability that the noise and inter-cell interference will make $\Re(z_k)$ negative. Since the expressions in Theorem 2 are fairly complicated, we also study the asymptotic behavior as $M \to \infty$.

Corollary 1: For large M, the complementary CDF $\Pr\{\Re(z_k) > b\}$ in (39) converges to

$$Q\left(\frac{b - \frac{\Gamma\left(M+\frac{1}{2}\right)}{\Gamma(M)}\sqrt{\lambda_1}}{\sqrt{\lambda_1^2\left(M - \left(\frac{\Gamma\left(M+\frac{1}{2}\right)}{\Gamma(M)}\right)^2\right) + \sigma^2 + \Upsilon_k + q\beta_k\tau_p - \lambda_1}}\right) \tag{40}$$

where the Q-function was defined in Theorem 2. In particular, for $\epsilon_k < \rho_k\beta_k\tau_p/2$, the probability becomes

$$\lim_{M\to\infty} \Pr\{\mathcal{R}_k\} = \begin{cases} 0, & \rho_k\beta_k\tau_p < \alpha_t/2 + \epsilon_k, \\ 1/2, & \rho_k\beta_k\tau_p = \alpha_t/2 + \epsilon_k, \\ 1, & \rho_k\beta_k\tau_p > \alpha_t/2 + \epsilon_k. \end{cases} \tag{41}$$

Proof: The proof is given in Appendix B. ∎

This corollary confirms that if one UE has a signal gain that is larger than the sum of all the others' signal gains (plus inter-cell interference), then this UE will repeat its pilot in Step 3 with probability one in the regime of large number of antennas. The bias parameter ϵ_k can be used to tune this condition; for example, at most one UE will transmit the pilot

[5]For all such bias terms, the estimate $\hat{\alpha}_{t,k} = \rho_k\beta_k\tau_p$ leads to a decision to repeat the pilot, which makes perfect sense since the UE believes that it is the only one that transmitted the pilot.
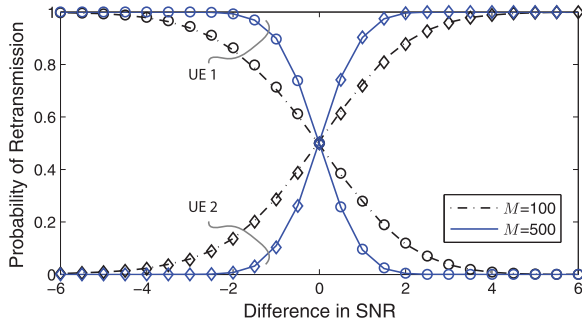
(a) Probability to repeat the pilot transmission in a two-UE collision
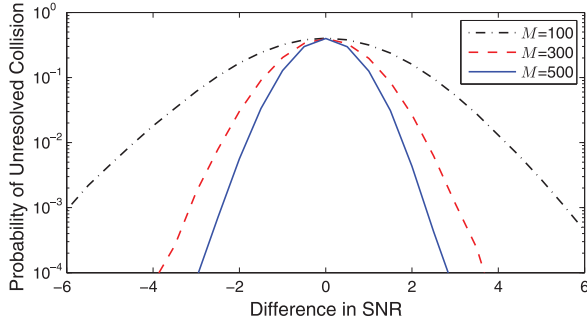
(b) Probability of having an unresolved two-UE collision

Fig. 6.   Two-UE collisions are studied where UE 1 has $\rho_1\beta_1\tau_p = 10\,\text{dB}$ and UE 2 has $\rho_2\beta_2\tau_p$ between 4 dB and 16 dB. The SNR difference is $\rho_2\beta_2\tau_p - \rho_1\beta_1\tau_p$. (a) The probabilities that each of the UEs repeat the pilot in Step 3. (b) The probability of having an unresolved collision.

in Step 3 when $\epsilon_k > 0$ for all k. At finite M there is always a risk for collisions and the pilot repetition decisions are correlated between the UEs, since they are based on partially the same sources of randomness (i.e., channel realizations and inter-cell interference). Since the expression for $\Pr\{\mathcal{R}_k\}$ in Theorem 2 is complicated, an exact computation of the probability to resolve collisions appears intractable. We will thus study this numerically instead.

C. Example: Resolving a Two-UE Collision

Let us consider a case where two UEs collide: $\mathcal{S}_t = \{1, 2\}$. The first UE has the fixed pilot SNR $\text{SNR}_1 = \rho_1\beta_1\tau_p = q\beta_1\tau_p = 10\,\text{dB}$, while the corresponding SNR_2 of the second UE is varied between 4 dB and 16 dB (with normalized noise variance $\sigma^2 = 1$). Fig. 6(a) shows the probability that the UEs repeat their pilot transmissions in Step 3, assuming $\epsilon_k = 0$ and either $M = 100$ or $M = 500$ BS antennas. The horizontal axis shows the SNR difference $\text{SNR}_2 - \text{SNR}_1$ between the UEs, which is between $-6\,\text{dB}$ and $+6\,\text{dB}$. The curves were generated by Monte-Carlo simulations, while the markers are computed using the closed-form expression in Theorem 2. The minor discrepancies are due to the finite number of Monte Carlo realizations. If there is an SNR difference of at least 3 dB, then the UE with the largest SNR is likely the only one to repeat the pilot. In contrast, both UEs repeat the pilot with equal probability when they have identical SNR. The transition between these cases is sharper the more antennas are used, which is in line with Corollary 1.

The probability of having an unresolved two-UE collision, $1 - P_{2,\text{resolved}}$ with $P_{2,\text{resolved}}$ defined in (21), is shown in Fig. 6(b) for $\epsilon_k = 0$ and $M \in \{100, 300, 500\}$. The proposed SUCRe protocol resolves almost all collisions when SNR_1 and SNR_2 are sufficiently different (e.g., more than 90% of the two-UE collisions when there is a 3 dB difference in SNR). The probability of having an unresolved collision drops rapidly when adding more antennas, except in the special case of $\text{SNR}_1 = \text{SNR}_2$ where it is stable at around 40%. If the decisions would have been independent between the UEs, then 50% of the collisions would be unresolved in this special case. Hence, the errors in the estimates $\hat{\alpha}_{t,1}^{\text{approx2}}$ and $\hat{\alpha}_{t,2}^{\text{approx2}}$ are correlated; when one UE experiences a small-scale channel realization much stronger than the average it will underestimate α_t, while the other UE is likely to overestimate α_t since it will believe that the other UE has a stronger average channel gain than it actually has.

This example shows that SNR differences are desirable when using the SUCRe protocol, which means that we should embrace rather than fully combat the pathloss variations that appear between UEs in cellular networks. We elaborate on this in the following remark.

Remark 4 (Uplink Pilot Power Control): Power control is used in conventional RA, such as LTE, to give the UEs equal conditions (i.e., the same signal gain $\rho_k\beta_k$). UEs at the cell edge transmit at full power, while cell-center UEs reduce their power. The SUCRe protocol prefers the opposite type of power control; each UE transmits at full power and the pathloss variations are exploited to resolve potential RA collisions. If the pathloss difference are larger than the dynamic range of the BS receiver, some fractional power control is needed to partially reduce the pathloss differences. In many practical cases it is very likely that the colliding UEs have pathloss differences (in terms of β_k) of at least 3 dB. In situations where this is not the case, the RA pilot powers ρ_k can be randomized around a nominal value to create a similar type of randomness where the colliding UEs are unlikely to have the same SNR. We will illustrate this numerically in the next section. Randomized power control can also increase the fairness (see Remark 3) by making the probability of being the strongest UE less dependent on β_k, which can be important in systems where random access is used for decentralized handover between cells. Ideal fairness can, in principle, be achieved by first applying conventional power control that makes $\rho_k\beta_k$ equal for all UEs and then generate random power variations around this nominal value.

IV. NUMERICAL RESULTS

In this section, we show numerically how the SUCRe protocol performs in cellular networks. The simulation results can be reproduced using the Matlab code that is available at https://github.com/emilbjornson/sucre-protocol. We consider the center cell of the hexagonal network depicted in Fig. 2 and take the activities in the six neighboring cells into account. The radius of each hexagon is 250 m, and the UEs are uniformly distributed in each cell at locations further than 25 m from the BS. The estimate $\hat{\alpha}_{t,k}^{\text{approx2}}$ of α_t is used in all the simulations.

A. Channel Propagation Models

We will compare three different channel models. The first one is uncorrelated Rayleigh fading, where $\mathbf{h}_k$ is distributed as in (22) for $k = 1, \ldots, K_0$. The second one is correlated Rayleigh fading with $\mathbf{h}_k \sim \mathcal{CN}(\mathbf{0}, \beta_k \mathbf{R}_k)$, where we consider a ULA at the BS modeled by the exponential correlation model with the correlation $r = 0.7$ between adjacent antennas [27]:

$$[\mathbf{R}_k]_{i,j} = r^{-|j-i|} e^{J \theta_k (j-i)} \tag{42}$$

where θ_k is the angle between UE k and BS 0. This represents a non-line-of-sight scenario with spatial correlation, meaning that the channel is statistically stronger in some spatial directions (determined by θ_k) than other directions.

The third model describes pure line-of-sight (LoS) propagation when the BS is equipped with a ULA with half-wavelength antenna spacing[6]:

$$\mathbf{h}_k = \sqrt{\beta_k} \left[1 \ e^{-J\pi \sin(\phi_k)} \ \ldots \ e^{-J\pi(M-1)\sin(\phi_k)} \right]^{\mathrm{T}}. \tag{43}$$

Note that this channel vector is deterministic, in contrast to the previous two models.

The pathloss is modeled based on the urban micro scenario in [28]. The Rayleigh fading cases have pathloss exponent 3.8 and shadow fading with log-normal distribution and standard deviation 10 dB, while the LoS case has pathloss exponent 2.5 and log-normal variations with standard deviation 4 dB. When a UE in a corner point of the cell transmits at full power the median of the SNR, $\rho_k \beta_k / \sigma^2$, is 0 dB in the non-LoS case and 33 dB in the LoS case. The BS and UEs transmit at the same constant power (i.e., $\rho_k = q$) giving the same SNR in both directions.

We will consider both cases when the adjacent cells are silent during the RA protocol and when they perform regular data transmission. In the latter case, we assume that there are ten active UEs in each of the neighboring cells and the propagation channels are modeled as uncorrelated Rayleigh fading (using the same power levels and pathloss models as above).[7] The average UL interference $\bar{\omega} = \mathbb{E}\{\|\mathbf{W} \frac{\boldsymbol{\psi}_t^*}{\|\boldsymbol{\psi}_t\|}\|^2 / M\}$, where the expectation is computed with respect to user locations and shadow fading realizations, is assumed to be known at the UE (it is the same for all UEs) and is subtracted from $\hat{\alpha}_{t,k}$ by setting $\epsilon_k = -\bar{\omega}/2$.

B. Probability to Resolve Collisions

We will now illustrate that the SUCRe protocol is capable of resolving collisions, even when the system is overloaded. We consider a scenario with $\tau_p = 10$ and $K_0 = 5000$ inactive UEs in the cell, where each UE accesses the network with a 0.5% probability in a given RA block. The number of UEs, $|\mathcal{S}_t|$, is distributed as illustrated in Fig. 7, which is obtained from the binomial distribution in (1) by conditioning on that $|\mathcal{S}_t| \geq 1$. This is an overloaded scenario in the sense that, on average, 2.5 UEs select each pilot, leading to collisions

[6]All elements in (43) should also be multiplied with a common phase-shift, but we neglect it here since it has no impact on the performance.

[7]We make sure that each BS has a hexagonal coverage area by only considering shadow fading realizations where each UE gets the highest signal gain from its serving BS.
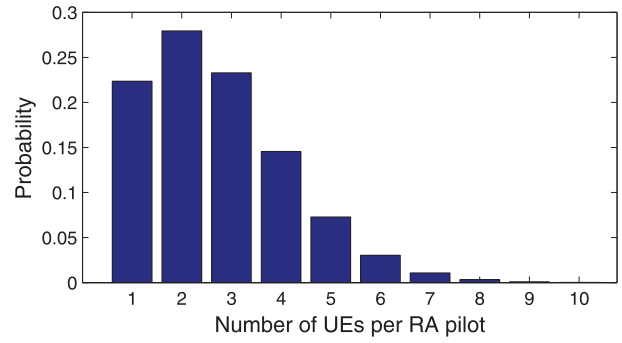

Fig. 7. Example of a distribution of the number of UEs that selects each RA pilot. It is used in Figs. 8 and 9.

at more than 75% of the pilots. Even collisions with 5 or 6 UEs occur frequently.

The probability to resolve collisions is defined as

$$P_{\text{resolved}} = \mathbb{E} \left\{ P_{|\mathcal{S}_t|,\text{resolved}} \big| |\mathcal{S}_t| \geq 1 \right\}$$

$$= \sum_{N=1}^{K_0} \frac{P_{N,\text{resolved}} \binom{K_0}{N} \left(\frac{P_a}{\tau_p} \right)^N \left(1 - \frac{P_a}{\tau_p} \right)^{K_0-N}}{1 - \left(1 - \frac{P_a}{\tau_p} \right)^{K_0}}, \tag{44}$$

where $P_{N,\text{resolved}}$ was defined in (20) and the expectation is with respect to $|\mathcal{S}_t|$. Fig. 8 shows this probability to resolve collisions as a function of the number of BS antennas, for the aforementioned three channel models. Fig. 8(a) considers inter-cell interference, while Fig. 8(b) neglects the interference (i.e., the adjacent cells are silent in the RA block).

The first observation from Fig. 8 is that the SUCRe protocol relies on the channel hardening and favorable propagation of Massive MIMO channels; P_{resolved} is 20-40% at $M = 1$, but increases steeply to 75-90% when having $M = 50$ antennas. The probability to resolve collisions continues to increase for $M \geq 50$, but at a slower pace. Uncorrelated Rayleigh fading gives better results than correlated Rayleigh fading, but the difference is small when M is large. The LoS model has slightly worse performance, because the pathloss differences are lower which makes it harder to appoint a strongest UE. However, since the cell-edge SNR is higher in the LoS case we can afford to create SNR differences by randomizing the UL pilot powers (as discussed in Remark 4). Fig. 8 also shows the LoS case when each UE reduces its pilot power with a random number between 0 dB and −30 dB, uniformly distributed in dB-scale. This case gives the highest performance among all the cases. Hence, the SUCRe protocol is well suited for both LoS and non-LoS channels.

As a baseline, we also consider a conventional protocol where pilot collisions are only handled by retransmission in later RA blocks. Specifically, using Fig. 1, all UEs deterministically repeat the pilots in Step 3 of the access protocol and proceed to the process of centralized contention resolution. Fig. 8 shows the probability to resolve collisions and we notice that the SUCRe protocol is able to admit roughly four times as many UEs per RA pilot than the baseline scheme. We also notice that inter-cell interference only causes a minor degradation in performance, except in the LoS case where performance is almost unaffected. The full-power LoS case performs similarly to the other channel models in this case.

(a) With interference from adjacent cells

(a) With interference from adjacent cells
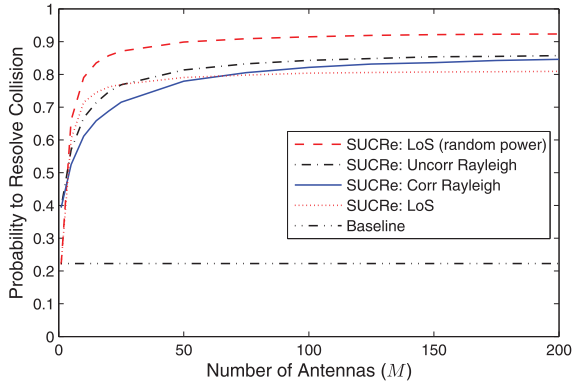
(b) Without interference from adjacent cells
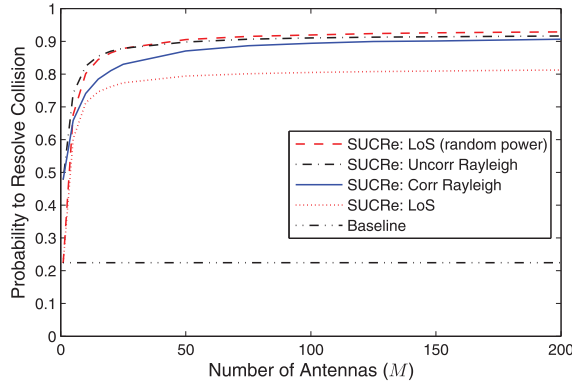
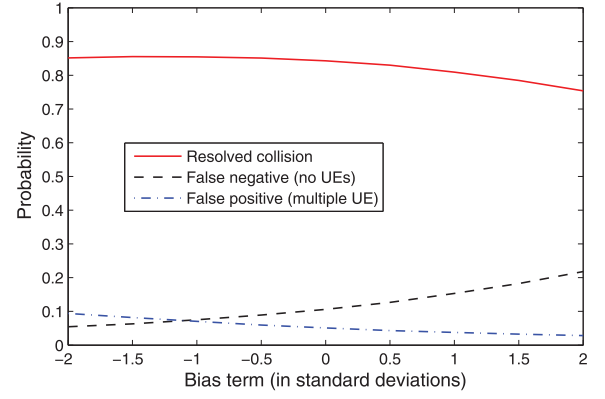(b) Without interference from adjacent cells

Fig. 8. Probability of resolving collisions, as a function of the number of BS antennas, in a highly loaded cellular network with or without inter-cell interference.

Fig. 9. Probability of resolving collisions, false negatives, and false positives for different bias terms in the decision rule, for uncorrelated Rayleigh fading channels.

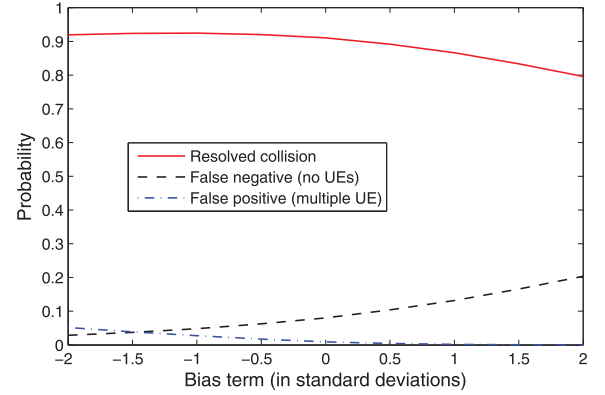### C. Tuning Probabilities Using Bias Term

Next, we consider the same scenario but focus on uncorrelated Rayleigh fading channels and study the probabilities of resolving collisions, false negatives, and false positives. We will demonstrate how the bias term ϵ_k can be utilized to tune the decisions by setting $\epsilon_k = \delta\beta_k/\sqrt{M} - \bar{\omega}/2$, which corresponds to adding δ standard deviations of $\|\mathbf{h}_k\|^2/M$ (around its mean value β_k). Fig. 9 shows the probabilities as a function of δ, for $M = 100$ and with or without inter-cell interference.

By subtracting one or two standard deviations from $\hat{\alpha}_{t,k}/2$ in the decision rule, we can encourage UE k to appoint itself the contention winner. This leads to higher probability of resolving collisions, at the cost of more false positives where multiple UEs repeat their pilot transmissions in Step 3. In contrast, by adding one or two standard deviations to $\hat{\alpha}_{t,k}/2$ in the decision rule, we can discourage UE k from appointing itself the contention winner and thereby push the probability of false positives towards zero—at the cost of resolving fewer collisions and having more false negatives where no UEs repeat their pilots in Step 3.

The probability of resolving a collision is naturally decreasing as the number of colliding users increases. Our simulation results reveal that the SUCRe protocol with $\delta = -1$ resolves 92% of the two-UE collisions, 82% of the five-UE collisions, and 71% of ten-UE collisions. Note that it is

unlikely to have more than a handful of colliding UEs per RA pilot, except when the network is extremely overloaded.

D. Average Number of RA Attempts in Crowded Scenarios

We have previously shown that the SUCRe protocol can resolve RA collisions, but the main purpose of an RA protocol is that every UE should be admitted to the data blocks after as few RA attempts as possible. We study this performance indicator in a scenario with $M = 100$, $\tau_p = 10$, and varying number of inactive UEs: $K_0 \in [100, 12000]$. Each UE decides to access the network with 0.1% probability. If it is not admitted immediately, then in the upcoming blocks the UE joins the SUCRe process that runs in that block with probability 0.5. If the UE has not succeeded after sending RA pilots in a total of 10 SUCRe processes (including the initial one), then it stops the transmission; that is, it considers that access has been denied by the network. Note that the procedure of joining the 9 additional SUCRe processes can be optimized according to the principles of splitting tree protocols [29], [30], but this optimization is outside the scope of this paper. We consider uncorrelated Rayleigh fading, cases with and without inter-cell interference, and we use the bias term $\epsilon_k = -\beta_k/\sqrt{M} - \bar{\omega}/2$. The SUCRe protocol is compared with the same baseline protocol as before (i.e., it only handles collisions by retransmission) with the addition that the UEs make 10 RA attempts at random occasions in the same way as the SUCRe protocol.
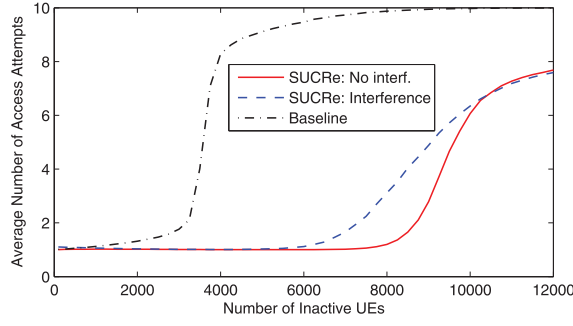
(a) Average number of RA attempts
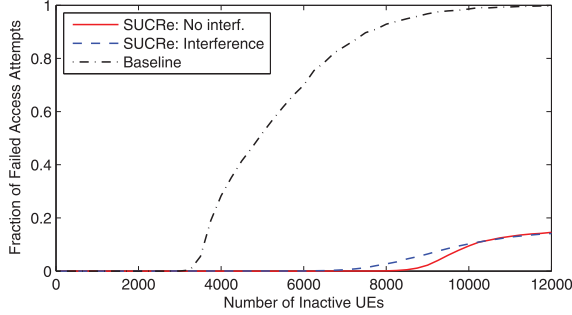
(b) Probability of failed RA attempt (more than 10 attempts)

Fig. 10. RA performance in a cellular network, where each UE accesses the network with 0.1% probability and sends 10 RA pilots before giving up. The SUCRe protocol can handle substantially higher user loads, K_0, than conventional methods.

Fig. 10(a) shows the average number of RA attempts that each UE makes, as a function of K_0, while Fig. 10(b) shows the fraction of UEs that fails to access the network (i.e., made 10 unsuccessful attempts). The SUCRe protocol can easily handle up to $K_0 = 6000$ under inter-cell interference and $K_0 = 8000$ when the adjacent cells are silent in the RA blocks. For larger values of K_0 around 0–15% of the UEs will fail to be admitted. Notice that at $K_0 = 10000$ there will on average be $K_0 \cdot 0.001/\tau_p = 1$ UE that selects each RA pilot, meaning that the network is fundamentally overloaded. Nevertheless, an astonishing 90% of the UEs can still access the network successfully, which matches well with the 90% probability of resolving collisions observed in Fig. 8. This behavior remains also for $K_0 > 10000$. In contrast, the baseline protocol requires more retransmissions when $K_0 < 3000$ and as K_0 increases in the range $K_0 > 3000$ the RA functionality gradually breaks down; at $K_0 = 10000$ only 1.5% of the UEs are successful in their RA attempts.

V. CONCLUSION

The pilot sequences are precious resources in Massive MIMO since they enable the BS to separate the UEs in the spatial domain. In future urban scenarios, the number of UEs that resides in a cell is much larger than the number of available pilots, thus the pilots need to be temporally allocated only to the UEs that have data to transmit or receive. The proposed SUCRe random access protocol provides an efficient way for UEs to request pilots for data transmission, and is well-suited for beyond-LTE Massive MIMO systems and crowded urban deployment scenarios. The protocol exploits the channel hardening and favorable propagation properties to enable distributed collision detection and resolution at the UEs, where the contender with the strongest signal gain is the one being admitted. The numerical results demonstrate that the SUCRe protocol can resolve around 90% of all collisions and that it is robust to inter-cell interference and choice of channel distribution. The protocol does not break down in overloaded situations, where more UEs request pilots than there are RA resources, but continues to admit a subset of the accessing UEs.

APPENDIX A
SOME USEFUL RESULTS

Lemma 2 (§8.328.2 in [31]): The Gamma function satisfies

$$\frac{\Gamma\left(M + \frac{1}{2}\right)}{\sqrt{M}\,\Gamma(M)} \to 1 \quad \text{as} \quad M \to \infty. \tag{45}$$

Lemma 3: For any non-negative integer m and real-valued A and B, we have

$$\int_0^\infty x^m e^{-(xA-B)^2} dx$$

$$= \begin{cases} \sum_{n=0}^{m} \binom{m}{n} \frac{B^{m-n}}{A^{m+1}} \frac{\Gamma\left(\frac{n+1}{2}\right) + (-1)^n \gamma\left(\frac{n+1}{2}, B^2\right)}{2}, & B \geq 0, \\ \sum_{n=0}^{m} \binom{m}{n} \frac{B^{m-n}}{A^{m+1}} \frac{\Gamma\left(\frac{n+1}{2}\right) - \gamma\left(\frac{n+1}{2}, B^2\right)}{2}, & B < 0. \end{cases} \tag{46}$$

Proof: By the change of variable $x = \frac{\tilde{x}+B}{A}$ we obtain

$$\int_0^\infty x^m e^{-(xA-B)^2} dx = \int_{-B}^\infty \left(\frac{\tilde{x}+B}{A}\right)^m \frac{e^{-\tilde{x}^2}}{A} d\tilde{x}$$

$$= \sum_{n=0}^{m} \binom{m}{n} \frac{B^{m-n}}{A^{m+1}} \int_{-B}^\infty \tilde{x}^n e^{-\tilde{x}^2} d\tilde{x}, \tag{47}$$

where the second equality follows from applying the binomial formula to $(\tilde{x}+B)^m$. If $B \geq 0$, the remaining integral is computed as

$$\int_{-B}^\infty \tilde{x}^n e^{-\tilde{x}^2} d\tilde{x} = \int_0^\infty \tilde{x}^n e^{-\tilde{x}^2} d\tilde{x} + (-1)^n \int_0^B \tilde{x}^n e^{-\tilde{x}^2} d\tilde{x}$$

$$= \frac{\Gamma\left(\frac{n+1}{2}\right) + (-1)^n \gamma\left(\frac{n+1}{2}, B^2\right)}{2} \tag{48}$$

by making the variable substitution $y = \tilde{x}^2$ and identifying incomplete gamma functions. Similarly, if $B < 0$ we have

$$\int_{-B}^\infty \tilde{x}^n e^{-\tilde{x}^2} d\tilde{x} = \int_0^\infty \tilde{x}^n e^{-\tilde{x}^2} d\tilde{x} - \int_0^{-B} \tilde{x}^n e^{-\tilde{x}^2} d\tilde{x}$$

$$= \frac{\Gamma\left(\frac{n+1}{2}\right) - \gamma\left(\frac{n+1}{2}, B^2\right)}{2}. \tag{49}$$

Substituting (48) or (49) into (47) yield the final result. ∎

APPENDIX B
COLLECTION OF PROOFS

Proof of Lemma 1: Suppose for a moment that $\mathcal{S}_t$ is known, then the MMSE estimator of $\mathbf{h}_k$ from the observation $\mathbf{y}_t$ at BS 0 is [32]

$$\hat{\mathbf{h}}_{k,\text{MMSE}} = \frac{\sqrt{\rho_k \tau_p}\,\beta_k}{\alpha_t + \sigma^2} \mathbf{y}. \tag{50}$$

The true channel can be expressed as $\mathbf{h}_k = \hat{\mathbf{h}}_{k,\mathrm{MMSE}} + \mathbf{e}_k$, where the estimate

$$\hat{\mathbf{h}}_{k,\mathrm{MMSE}} \sim \mathcal{CN}\left(\mathbf{0}, \frac{\rho_k \tau_p \beta_k^2}{\alpha_t + \sigma^2}\mathbf{I}_M\right) \tag{51}$$

is independent from the estimation error

$$\mathbf{e}_k \sim \mathcal{CN}\left(\mathbf{0}, \left(\beta_k - \frac{\rho_k \tau_p \beta_k^2}{\alpha_t + \sigma^2}\right)\mathbf{I}_M\right). \tag{52}$$

The BS does not know $\mathcal{S}_t$, but we notice that

$$\frac{\hat{\mathbf{h}}_{k,\mathrm{MMSE}}^*}{\|\hat{\mathbf{h}}_{k,\mathrm{MMSE}}\|} = \frac{\mathbf{y}^*}{\|\mathbf{y}\|}, \tag{53}$$

thus the received signal in (13) can be rewritten as

$$\begin{aligned}
z_k &= \sqrt{q\tau_p}\,\mathbf{h}_k^{\mathrm{T}}\frac{\hat{\mathbf{h}}_{k,\mathrm{MMSE}}^*}{\|\hat{\mathbf{h}}_{k,\mathrm{MMSE}}\|} + \boldsymbol{v}_k^{\mathrm{T}}\frac{\boldsymbol{\phi}_t^*}{\|\boldsymbol{\phi}_t\|} + \eta_k \\
&= \underbrace{\sqrt{q\tau_p}\|\hat{\mathbf{h}}_{k,\mathrm{MMSE}}\|}_{=g_k} \\
&\quad + \underbrace{\sqrt{q\tau_p}\frac{\mathbf{e}_k^{\mathrm{T}}\hat{\mathbf{h}}_{k,\mathrm{MMSE}}^*}{\|\hat{\mathbf{h}}_{k,\mathrm{MMSE}}\|} + \boldsymbol{v}_k^{\mathrm{T}}\frac{\boldsymbol{\phi}_t^*}{\|\boldsymbol{\phi}_t\|} + \eta_k}_{=v_k},
\end{aligned} \tag{54}$$

where we call the two terms g_k and v_k. We notice that

$$\frac{\mathbf{e}_k^{\mathrm{T}}\hat{\mathbf{h}}_{k,\mathrm{MMSE}}^*}{\|\hat{\mathbf{h}}_{k,\mathrm{MMSE}}\|} \sim \mathcal{CN}\left(0, \beta_k - \frac{\rho_k \tau_p \beta_k^2}{\alpha_t + \sigma^2}\right) \tag{55}$$

since $\mathbf{e}_k$ is independent from the channel estimate and $\frac{\hat{\mathbf{h}}_{k,\mathrm{MMSE}}}{\|\hat{\mathbf{h}}_{k,\mathrm{MMSE}}\|}$ is uniformly distributed over the unit sphere in $\mathbb{C}^M$. Hence, g_k and v_k are independent random variables. In addition, v_k is the sum of three independent complex Gaussian variables which have zero mean and the total variance as stated in the lemma. Finally, we notice that g_k^2 is the sum of squares of $2M$ independent Gaussian variables with zero mean and variance $\frac{1}{2}\frac{\rho_k q \beta_k^2 \tau_p^2}{\alpha_t + \sigma^2}$, thus g_k has a scaled chi-distribution with $2M$ degrees of freedom as stated in the lemma.

Proof of Theorem 1: The ML estimator is defined as

$$\hat{\alpha}_{t,k}^{\star} = \arg\max_{\alpha}\ f\left(z_{k,\Re}, z_{k,\Im}|\alpha\right) \tag{56}$$

where $f\left(z_{k,\Re}, z_{k,\Im}|\alpha\right)$ is the joint PDF. Since the UE knows that $\rho_k \beta_k \tau_p$, it is sufficient to search for $\alpha \geq \rho_k \beta_k \tau_p$. Notice that $z_{k,\Re} = g_k + \Re(v_k)$ and $z_{k,\Im} = \Im(v_k)$ are independent since $v_k \sim \mathcal{CN}(0, \lambda_2)$ has independent real and imaginary parts that are distributed as $\mathcal{N}(0, \lambda_2/2)$. Hence,

$$f\left(z_{k,\Re}, z_{k,\Im}|\alpha\right) = f_1\left(z_{k,\Re}|\alpha\right) f_2\left(z_{k,\Im}|\alpha\right) \tag{57}$$

where $f_2\left(z_{k,\Im}|\alpha\right)$ in (31) is the PDF of $\Im(v_k)$. It remains to compute the PDF $f_1\left(z_{k,\Re}|\alpha\right)$ of $z_{k,\Re}$, which is a convolution of the PDFs of g_k and $\Re(v_k)$:

$$\begin{aligned}
&f_1\left(z_{k,\Re}|\alpha\right) \\
&= \int_0^{\infty} \frac{2x^{2M-1}e^{-x^2/\lambda_1}}{\Gamma(M)\lambda_1^M}\frac{e^{-(z_{k,\Re}-x)^2/\lambda_2}}{\sqrt{\pi\lambda_2}}dx
\end{aligned}$$

$$\begin{aligned}
&= \frac{2e^{-z_{k,\Re}^2/\lambda_2}}{\Gamma(M)\lambda_1^M\sqrt{\pi\lambda_2}}\int_0^{\infty}x^{2M-1}e^{-x^2(\frac{1}{\lambda_1}+\frac{1}{\lambda_2})+x\frac{2z_{k,\Re}}{\lambda_2}}dx \\
&= \frac{2e^{-\frac{(z_{k,\Re})^2}{\lambda_2}\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)}}{\Gamma(M)\lambda_1^M\sqrt{\pi\lambda_2}}\int_0^{\infty}x^{2M-1}e^{-(xA-B)^2}dx
\end{aligned} \tag{58}$$

where $A = \sqrt{\frac{1}{\lambda_1}+\frac{1}{\lambda_2}}$ and $B = \frac{z_{k,\Re}}{\lambda_2 A}$. The final expression for f_1 in (57) is obtained by computing the integral in (58) using Lemma 3 with $m = 2M - 1$.

Proof of Theorem 2: The probability of repeating the pilot in Step 3, based on (18), is

$$\Pr\{\mathcal{R}_k\} = \Pr\left\{2\rho_k\beta_k\tau_p > \hat{\alpha}_{t,k}^{\mathrm{approx2}} + 2\epsilon_k\right\}. \tag{59}$$

Notice that whenever $\epsilon_k < \rho_k\beta_k\tau_p/2$, $\hat{\alpha}_{t,k}^{\mathrm{approx2}} = \rho_k\beta_k\tau_p$ will always lead to pilot repetition, thus we can neglect the maximum-operator in (36) and write

$$\begin{aligned}
\Pr\{\mathcal{R}_k\} &= \Pr\left\{2\rho_k\beta_k\tau_p > C_M^2\frac{q\rho_k\beta_k^2\tau_p^2}{(\Re(z_k))^2} - \sigma^2 + 2\epsilon_k\right\} \\
&= \Pr\left\{(\Re(z_k))^2 > \zeta_k\right\}
\end{aligned} \tag{60}$$

where we use the notation $C_M = \Gamma\left(M+\frac{1}{2}\right)/\Gamma(M)$ and the second equality follows from rearranging the term and identifying ζ_k from (38). Since $\Re(z_k)$ can be negative, we rewrite (60) as

$$\begin{aligned}
\Pr\{\mathcal{R}_k\} &= \Pr\left\{\Re(z_k) > \sqrt{\zeta_k}\right\} + \Pr\left\{\Re(z_k) < -\sqrt{\zeta_k}\right\} \\
&= 1 - \Pr\left\{\Re(z_k) \leq \sqrt{\zeta_k}\right\} + \Pr\left\{\Re(z_k) \leq -\sqrt{\zeta_k}\right\}.
\end{aligned} \tag{61}$$

which utilizes the fact that $\Pr\{\Re(z_k) < -\sqrt{\zeta_k}\} = \Pr\{\Re(z_k) \leq -\sqrt{\zeta_k}\}$. It remains to compute the CDF $\Pr\{\Re(z_k) \leq d\}$ for an arbitrary d, which is the convolution of the CDF of g_k and the PDF of $\Re(v_k)$:

$$\begin{aligned}
&\Pr\{\Re(z_k) \leq b\} \\
&= \int_0^{\infty}\frac{\gamma\left(M, \frac{x^2}{2\lambda_1}\right)}{\Gamma(M)}\frac{e^{-(b-x)^2/\lambda_2}}{\sqrt{\pi\lambda_2}}dx \\
&= \int_0^{\infty}\frac{e^{-(b-x)^2/\lambda_2}}{\sqrt{\pi\lambda_2}}dx \\
&\quad - \int_0^{\infty}e^{-x^2/\lambda_1}\sum_{k=0}^{M-1}\frac{x^{2k}}{\lambda_1^k\Gamma(k+1)}\frac{e^{-(b-x)^2/\lambda_2}}{\sqrt{\pi\lambda_2}}dx
\end{aligned} \tag{62}$$

by utilizing the definition of the incomplete gamma function (see Footnote 3). The first integral in (62) is over a Gaussian PDF and identified as $Q(-b\sqrt{2/\lambda_2})$. The second integral can be rewritten as

$$\begin{aligned}
&\sum_{k=0}^{M-1}\frac{e^{-b^2/\lambda_2}}{\Gamma(k+1)\lambda_1^k\sqrt{\pi\lambda_2}}\int_0^{\infty}x^{2k}e^{-x^2(\frac{1}{\lambda_1}+\frac{1}{\lambda_2})+x\frac{2b}{\lambda_2}}dx \\
&= \sum_{k=0}^{M-1}\frac{e^{-\frac{b^2}{\lambda_2}\left(1-\frac{\lambda_1}{\lambda_1+\lambda_2}\right)}}{\Gamma(k+1)\lambda_1^k\sqrt{\pi\lambda_2}}\int_0^{\infty}x^{2k}e^{-(xA-B)^2}dx
\end{aligned} \tag{63}$$

where $A = \sqrt{\frac{1}{\lambda_1}+\frac{1}{\lambda_2}}$ and $B = \frac{b}{\lambda_2 A}$. The final expression in (39) is obtained from (62)–(63) by computing the remaining integral using Lemma 3 with $m = 2k$.

Proof of Corollary 1: The Lindeberg-Lévy central limit theorem implies that g_k converges to $\mathcal{N}(\sqrt{\lambda_1} C_M, \lambda_1^2(M - C_M^2))$ in distribution as $M \to \infty$, where $C_M = \Gamma\left(M + \frac{1}{2}\right) / \Gamma(M)$. Since $\Re(z_k) = g_k + \Re(\nu_k)$ converges to the sum of two independent Gaussian variables (recall Lemma 1), the CDF is obtained by (40).

Notice that (37) can be expressed as

$$\Pr\{\mathcal{R}_k\} = 1 + \Pr\{\Re(z_k) > \sqrt{\zeta_k}\}) - \Pr\{\Re(z_k) > -\sqrt{\zeta_k}\}. \quad (64)$$

Since ζ_k in (38) is positive, $\Pr\{\Re(z_k) > -\sqrt{\zeta_k}\} \to Q(-\infty) = 1$ as $M \to \infty$. Similarly, we notice that

$$\Pr\{\Re(z_k) > \sqrt{\zeta_k}\}) \to \begin{cases} 1, & \zeta_k < C_M^2 \lambda_1, \\ 1/2, & \zeta_k = C_M^2 \lambda_1, \\ 0, & \zeta_k > C_M^2 \lambda_1. \end{cases} \quad (65)$$

The asymptotic probabilities in (41) follows directly from these results.

REFERENCES

[1] "Ericsson mobility report," Ericsson, Stockholm, Sweden, Tech. Rep., Nov. 2015. [Online]. Available: http://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf

[2] M. Fallgren, B. Timus, *D1.1: Scenarios, Requirements and KPIs for 5G Mobile and Wireless System*, document ICT-317669-METIS, 2013.

[3] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[4] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[5] E. Björnson, E. G. Larsson, and M. Debbah, "Massive MIMO for maximal spectral efficiency: How many users and pilots should be allocated?" *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 1293–1308, Feb. 2016.

[6] E. Björnson, E. G. Larsson, and T. L. Marzetta, "Massive MIMO: Ten myths and one critical question," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 114–123, Feb. 2016.

[7] J. Vieira *et al.*, "A flexible 100-antenna testbed for massive MIMO," in *Proc. IEEE Globecom Workshop Massive (MIMO)*, Dec. 2014, pp. 287–293.

[8] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.

[9] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving 'massive MIMO' spectral efficiency with a not-so-large number of antennas," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3226–3239, Sep. 2012.

[10] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[11] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[12] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, Feb. 2013.

[13] A. Adhikary, J. Nam, J. Y. Ahn, and G. Caire, "Joint spatial division and multiplexing—The large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct. 2013.

[14] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.

[15] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014.

[16] E. Björnson and E. G. Larsson, "Three practical aspects of massive MIMO: Intermittent user activity, pilot synchronism, and asymmetric deployment," in *Proc. IEEE Globecom Workshops*, Dec. 2015, pp. 1–6.

[17] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[18] N. K. Pratas, H. Thomsen, C. Stefanovic, and P. Popovski, "Code-expanded random access for machine-type communications," in *Proc. IEEE Globecom Workshops*, Sep. 2012, pp. 1681–1686.

[19] J. H. Sørensen, E. de Carvalho, and P. Popovski, "Massive MIMO for crowd scenarios: A solution based on random access," in *Proc. IEEE Globecom Workshops*, Jun. 2014, pp. 352–357.

[20] E. de Carvalho, E. Björnson, E. G. Larsson, and P. Popovski, "Random access for massive MIMO systems with intra-cell pilot contamination," in *Proc. IEEE ICASSP*, Mar. 2016, pp. 3361–3365.

[21] L. Sanguinetti, A. A. D'Amico, M. Morelli, and M. Debbah, "Random access in uplink massive MIMO systems: How to exploit asynchronicity and excess antennas," in *Proc. IEEE GLOBECOM*, May 2016, pp. 1–5.

[22] E. Paolini, G. Liva, and M. Chiani, "High throughput random access via codes on graphs: Coded slotted ALOHA," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sep. 2011, pp. 1–6.

[23] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: How coding theory helps to build random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144–150, Jun. 2015.

[24] E. Björnson, E. de Carvalho, E. G. Larsson, and P. Popovski, "Random access protocol for massive MIMO: Strongest-user collision resolution (SUCR)," in *Proc. IEEE ICC*, May 2016, pp. 1–6.

[25] H. Ngo, E. Larsson, and T. Marzetta, "Aspects of favorable propagation in massive MIMO," in *Proc. EUSIPCO*, Sep. 2014, pp. 76–80.

[26] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[27] S. L. Loyka, "Channel capacity of MIMO architecture using the exponential correlation matrix," *IEEE Commun. Lett.*, vol. 5, no. 9, pp. 369–371, Sep. 2001.

[28] *Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations (Release 13)*, document TR 25.996, 3GPP, Dec. 2015.

[29] J. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. Inf. Theory*, vol. IT-25, no. 5, pp. 505–515, Sep. 1979.

[30] J. H. Sørensen, C. Stefanović, and P. Popovski, "Coded splitting tree protocols," in *Proc. IEEE ISIT*, Sep. 2013, pp. 2860–2864.

[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. San Diego, CA, USA: Academic, 1980.

[32] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

Elisabeth de Carvalho received the Ph.D. degree in electrical engineering from Telecom ParisTech, France. She was a Post-Doctoral Fellow with Stanford University, Standford, CA, USA, and then with industry in the field of DSL and wireless LAN. Since 2005, she has been an Associate Professor with Aalborg University, where she has led several research projects in wireless communications. She co-authored the text book *A Practical Guide to the MIMO Radio Channel*. Her main expertise is in signal processing for MIMO communications, with recent focus on massive MIMO, including channel measurements, channel modeling, beamforming, and protocol aspects.

Jesper H. Sørensen (M'10) received the B.Sc. degree in electrical engineering, the M.Sc. degree (*cum laude*), and the Ph.D. degree in wireless communications from Aalborg University, Denmark, in 2007, 2009, and 2012, respectively. In 2011, he was an Intern with Mitsubishi Electric Research Laboratories. From 2012 to 2016, he was a Post-Doctoral Researcher with Aalborg University, where he is currently an Assistant Professor. His work is in the areas of channel coding, multimedia communication, and protocol design.

Erik G. Larsson received the Ph.D. degree from Uppsala University, Sweden, in 2002.

He was with the KTH Royal Institute of Technology, Stockholm, Sweden, the University of Florida, USA, the George Washington University, Washington, DC, USA, and with Ericsson Research, Sweden. In 2015, he was a Visiting Fellow with Princeton University, Princeton, NJ, USA, for four months. He is currently a Professor of Communication Systems with Linköping University, Linköping, Sweden. He co-authored the two textbooks *Space-Time Block Coding for Wireless Communications* (Cambridge Univ. Press, 2003) and *Fundamentals of Massive MIMO* (Cambridge Univ. Press, 2016). He holds 16 issued and many pending patents on wireless technology. His main professional interests are within the areas of wireless communications and signal processing. He has co-authored some 130 journal papers on these topics.

Dr. Larsson was Associate Editor for, among others, the *IEEE Transactions on Communications* (2010-2014) and the *IEEE Transactions on Signal Processing* (2006-2010). From 2015 to 2016 he served as chair of the IEEE Signal Processing Society SPCOM technical committee, and in 2017 he is the past chair of this committee. From 2014 to 2015 he served as chair of the steering committee for the *IEEE Wireless Communications Letters*. He was the General Chair of the Asilomar Conference on Signals, Systems and Computers in 2015, and its Technical Chair in 2012. He is a member of the IEEE Signal Processing Society Awards Board during 2017–2019.

He received the IEEE Signal Processing Magazine Best Column Award twice, in 2012 and 2014, and the IEEE ComSoc Stephen O. Rice Prize in Communications Theory in 2015.

Petar Popovski (S'97–M'04–SM'10–F'16) received the Dipl.-Ing. degree in electrical engineering and the Magister Ing. degree in communication engineering from Sts. Cyril and Methodius University, Skopje, Macedonia, in 1997 and 2000, respectively, and the Ph.D. degree from Aalborg University, Denmark, in 2004. He is currently a Professor in wireless communications with Aalborg University. He has over 270 publications in journals, conference proceedings, and books. He holds over 30 patents and patent applications. His research interests are in the areas of wireless communication and networking, and communication/information theory. He received the Consolidator Grant from the European Research Council. He was a recipient of the Elite Researcher Award in Denmark in 2016. From 2012 to 2014, he served as the Chair of the IEEE ComSoc Emerging Technology Committee on Smart Grid Communications. He is a Steering Committee Member of the IEEE Internet of Things Journal and a Steering Committee Member of the IEEE SmartGridComm. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and an Area Editor of the IEEE TRANSACTIONS WIRELESS COMMUNICATIONS.