Coordinated Pilot Transmissions for Detecting the Signal Sparsity Level in Massive IoT Networks

Onel L. A. López, *Member, IEEE,* Glauber Brante, *Senior Member, IEEE,* Richard D. Souza, *Senior Member, IEEE,* Markku Juntti, *Fellow, IEEE,* and Matti Latva-aho, *Fellow, IEEE*

***Abstract*—Grant-free protocols exploiting compressed sensing multi-user detection (MUD) are appealing for solving the random access problem in massive Internet of Things (IoT) networks with sporadic device activity. Such protocols would greatly benefit from prior deterministic knowledge of the sparsity level, i.e., the instantaneous number of simultaneously active devices K. Aiming at this, herein we introduce a framework relying on coordinated pilot transmissions (CPTs) for detecting K. Specifically, the proposed CPT mechanism includes a downlink (DL) phase for channel state information acquisition that resolves fading uncertainty in the uplink (UL) transmission phase using shared UL pilot symbols for channel compensation. We propose a signal sparsity level detector and analytically assess its accuracy when network channels are subject to Rayleigh fading. We show that the variance of the estimator increases with K, and its distribution approximates that of the sum of a Student's t and Gaussian random variable. The numerical results evince the need for carefully configuring the duration of the DL and UL phases. Indeed, we show that relatively short DL phases are preferable in highly sparse networks given the total CPT duration is fixed. Finally, we discuss and exemplify with some early results the potential of the proposed CPT framework for MUD, and highlight relevant research directions.**

***Index Terms*—massive IoT, compressed sensing, multi-user detection, signal sparsity level, grant-free random access**

I. Introduction

The number of Internet of Things (IoT) devices is exponentially growing driven by the need to turn our homes, vehicles, entertainment, health, work, industries, and social/community services into smart, autonomous, sustainable, interactive, and intelligent environments [1]–[5]. The massive machine-type communication (mMTC) paradigm aims to address the corresponding connectivity challenges, which are intertwined with the unique features of massive IoT setups, specifically [1]–[4]: i) sporadic transmissions, i.e., an unknown/random subset of machine-type communication devices, called simply devices in the sequel, is active at a given time instant; ii) short-packet communications dominated by uplink (UL) traffic; and iii) energy-limited communications/operation. The third feature evinces the need for energy-efficient communication/operation protocols and, in many cases, battery-free operation [6], [7]. Meanwhile, all features, in particular the first two, call for novel multiple-access mechanisms [3]–[5], to which our work here contributes.

Grant-free multiple-access protocols are particularly attractive for mMTC since they [1]–[3]: i) promote efficient spectrum utilization as each device is not assigned a dedicated transmission resource block, ii) reduce signaling overhead, and iii) improve energy efficiency of the devices. Note that due to the massiveness of the network, it is impossible to assign orthogonal pilot sequences/preambles to the devices, thus, motivating the need for grant-free non-orthogonal multiple access protocols. However, a key challenge here lies in efficiently identifying the set of sporadically active, non-orthogonally coexisting, devices and their data, for which collision resolution mechanisms are required [3], [4].

We can distinguish two basic types of collisions: *hard* and *soft*. The former occurs when exactly the same preamble is used simultaneously by several active devices. In contrast, the latter occurs when active devices use different non-orthogonal preambles, as they interfere to some extent with each other. The probability of hard/soft collisions increases/decreases as the number of available preambles reduces. Since hard collisions are difficult to resolve without relying on sufficiently orthogonal channel subspaces [8]–[10] and/or additional communication overhead, increasing the pool of non-orthogonal preambles (thus, favoring the occurrences of soft instead of hard collisions) is usually recommended in practice [2]–[4]. A promising class of soft collision resolution methods, known as compressed sensing (CS) techniques, have been considered for multi-user detection (MUD) in mMTC [11], [12]. Note that MUD may include both user activity detection (UAD) and data detection, jointly or separately. However, in the following and to simplify our exposition, we refer by MUD either to i) UAD alone, in the case that data detection is implemented separately, or ii) both UAD and data detection, in the case they are implemented jointly.

A. Related Work

CS-MUD is usually based on regularization, greedy, message-passing (MP), and/or artificial intelligence (AI) techniques.

1) *Regularized MUD* relies on transforming the highly non-convex CS-MUD problem to convex via regularization and iterative procedures. For instance, Zhu and Giannakis [13]

O. López, M. Juntti, and M. Latva-aho are with the Centre for Wireless Communications University of Oulu, Finland, e-mails: {Onel.AlcarazLopez, Markku.Juntti, Matti.Latva-aho}@oulu.fi. G. Brante is with the Department of Electrotechnics, Federal University of Technology PR, Curitiba, Brazil, e-mail: gbrante@utfpr.edu.br. R. D. Souza is with the Electrical and Electronics Engineering Department, Federal University of Santa Catarina (UFSC), Florianópolis, Brazil, e-mail: richard.demo@ufsc.br.

This work has been supported by the Research Council of Finland (former Academy of Finland) 6G Flagship Programme (Grant Number: 346208), the Finnish Foundation for Technology Promotion, and in Brazil by the National Council for Scientific and Technological Development, and the Network for Education and Research (RNP/MCTIC 01245.020548/2021-07).

proposed a ridge detector and a least absolute shrinkage and selection operator detector, which directly regularize the original CS-MUD problem based on l_2- and l_1-norm, respectively. Later, some sparsity-aware successive interference cancellation regularization techniques were proposed in [14], [15] aiming at lowering the detection complexity by sequentially recovering the transmitted symbols. Meanwhile, Renna and Lamare [16] incorporated a l_1- norm regularization into an iteratively updated linear minimum mean square error filter and a constellation-list scheme to enable sparse detection. Moreover, joint user identification and channel estimation approaches using the alternating direction method of multipliers (ADMM) were proposed in [17]–[19]. Finally, Gao *et al.* [20] proposed a low complexity coordinate descent mechanism for the CS-MUD problem.

2) *Greedy MUD* has low complexity and often only requires appropriate termination tuning of the transmitted signal/vector reconstruction. Schepker and Dekorsy [21] applied for the first time orthogonal least squares and orthogonal matching pursuit (OMP) greedy algorithms to a sparse mMTC scenario. Since the latter outperforms the former, the latest research on greedy MUD has focused mainly on OMP-based algorithms. For instance, Schepker *et al.* [22] proposed group OMP leveraging channel decoders for greater performance, while Xiong *et al.* [23] proposed a detection-based OMP algorithm that, unlike conventional OMP, does not rely on prior knowledge of the signal/device sparsity (the number of active devices). Specifically, it runs a binary hypothesis test on the residual vector of OMP at each iteration, while stopping when there is no signal component in the residual vector. Meanwhile, a noise-robust greedy algorithm exploiting *a posteriori* probability ratios for every index of sparse input signals is designed in [24]. Lee and Yu [25] leveraged *a priori* information on the activation probability of each device to improve the performance of several greedy MUD schemes in mMTC, and showed that they are robust against prior information inaccuracy. Finally, Xiao *et al.* [26] proposed a MUD mechanism exploiting backward signal sparsity estimation. The latter is implemented by modifying the classical sparsity adaptive matching pursuit algorithm [27] to deal with data length diversity coming from the exploitation of repeating and spreading sequences.

3) *MP-based MUD* constitutes a class of algorithms that exploit factor graphs, thus the *a posteriori* distribution of the signal to be reconstructed. In practice, due to the large-scale nature of the access problem in mMTC, the usual approach is to adopt/design approximate MP (AMP) algorithms relying on iterative thresholding, which also allows analytic performance characterization via the so-called state evolution. For instance, Chen *et al.* [28] derived efficient denoisers for AMP depending on whether the large-scale component of the channel fading is known. Senel and Larsson [29] analyzed and proposed algorithmic enhancements for coherent and non-coherent MUD based on AMP. Meanwhile, Ke *et al.* [30] designed non-orthogonal pseudorandom pilots for massive UL broadband access. They formulated active user detection and channel estimation as a generalized multiple measurement vector CS problem and solved it via a generalized multiple measurement vector AMP algorithm. The suitability of AMP for joint device activity detection and channel estimation of devices coexisting with mobile broadband services is assessed and promoted in [31]. Wang *et al.* [32] designed an AMP algorithm that exploits the temporal activation correlation of the devices, and showed the achievable performance gains. Renna and Lamare [33] proposed the so-called bilinear message-scheduling generalized AMP, which uses the channel decoder's beliefs to refine activity detection and data decoding. An AMP-aided CSI estimator and MUD is proposed in [34], where the authors use a multi-state Markov chain-based transmission model to characterize the diverse time-varying traffic demands of the users. Finally, Ke *et al.* [35] proposed an AMP-based unified semi-blind detection framework for grant-free sourced and unsourced random access aiming to facilitate massive ultra-reliable low-latency (URLLC) in massive multiple-input multiple-output (MIMO) systems.

4) *AI-based MUD* leads to direct detection decisions as the detection parameters are learned and configured on the go, thus, avoiding empirical parameter tuning. Deep learning is the most commonly used AI tool for solving the CS-MUD problem [36]. Some examples of deep learning-based MUD can be found in [37]–[40]. Specifically, Bai *et al.* [37] proposed a fast data-driven algorithm for CS-MUD in mMTC relying on a novel block restrictive activation nonlinear unit that nicely captures the system sparsity. Meanwhile, Cui *et al.* [38] designed two model-driven approaches, which effectively utilize features of sparsity patterns in designing common measurement matrices and adjusting the state-of-the-art detectors/decoders. Interestingly, the optimum depth, i.e., the number of layers, to be configured in a deep neural network varies according to the sparsity statistics, which motivated the work in [39]. Therein, the authors proposed to autonomously/dynamically update the number of layers in the inference phase by introducing an extra halting score at each layer. Hanxiao *et al.* [40] proposed a deep learning approach consisting of a preamble detection neural network for a first tentative/rough MUD followed by a data detection neural network exploiting the information data signals to refine MUD. Finally, AI-based MUD may also leverage Bayesian learning [41]–[44]. Indeed, Zhang *et al.* [41] developed two CS-MUD Bayesian inference algorithms exploiting sparse prior information of the estimated channel vector. Similar approaches, but also exploiting the correlation of user activity over successive access slots, are proposed in [42]. Meanwhile, Marata *et al.* [43] proposed a unified framework for non-coherent and coherent mMTC and enhanced mobile broadband data transmissions, respectively, including a proper pilot design. MUD for clustered MTC is explored in [44] by utilizing the approximation error method to account for errors in the sensing matrix and likelihood function. In addition to Bayesian learning, the works in [43] and [44] also assessed the system performance under regularized, greedy, and MP-based MUD algorithms.

B. Contributions

In general, the state-of-the-art research on CS-MUD either assumes that i) signal sparsity level is known and exploited

for MUD, e.g., [13]–[19], or ii) signal sparsity level detection is a sub-product or stage of MUD, e.g., [20]–[35], [37]–[44]. In the first case, there have been no direct answers on how the sparsity level could be accurately known in advance to MUD in mMTC,[1] which makes the mechanisms proposed in [13]–[19] impractical so far. In the second case, the sparsity-level information is not required. However, having and exploiting this information would certainly improve the MUD performance. Specifically, the iterative mechanisms proposed in [20]–[35], [41]–[44] face the challenge of setting an appropriate stopping criterion, although the proposals in [26], [27] are more resilient in this regard, since the signal sparsity level is implicitly and iteratively estimated together with MUD. Moreover note that the deep learning-based mechanisms proposed in [37], [38], [40] have a fixed depth in terms of the number of layers and do not adapt well to highly-varying sparsity levels. Interestingly, the depth could also be learned [39], but this introduces further non-linearity into the system. Although such an approach leads to accuracy improvements with respect to state-of-the-art MUD based on deep learning, it is also more complex.

From the discussion above, and as highlighted in [12], we can conclude that the CS-MUD mechanisms can all significantly benefit from a (sufficiently) deterministic prior on the sparsity level, which is specifically our aim here. Information on the sparsity level enables the application of the MUD solutions in [13]–[19], while potentially making those in [20]–[35], [37]–[44] more easily configurable and accurate.

We consider an mMTC deployment under quasi-static fading, where K random devices become active and aim to communicate with a coordinator. Our main contributions are three-fold:

- We introduce a coordinated pilot transmission (CPT) framework for detecting[2] the signal sparsity level, K, in time division duplex (TDD) systems and to be implemented prior to the MUD. Specifically, the CPT mechanism consists of a downlink (DL) broadcast transmission using N_1 symbols for the purpose of channel state information (CSI) estimation, and an UL transmission with channel inversion (power and phase) control using N_2 shared symbols to resolve the fading uncertainty at the coordinator. Note that the use of shared pilot symbols is a key innovation here. After this, the signal sparsity level, K, is detected based on the signal received at the coordinator by performing a relaxed (real-domain) estimation followed by a rounding-to-the-nearest operation.
- We assess the performance of the proposed CPT mechanism and signal sparsity level estimator in Rayleigh fading channel conditions. Specifically, we characterize analytically the permissible maximum power and average power consumption of the devices, and the probability that an active device cannot transmit due to insufficient power to compensate for the channel losses. Moreover, we demonstrate and corroborate numerically that the estimator's variance increases linearly with K and that its distribution matches approximately that of the sum of a Student's t and a Gaussian random variable. Moreover, we provide a semi-closed-form approximation for the detection success probability under the proposed signal sparsity level estimator, which is valuable for system design/optimization purposes.
- We show that the attainable accuracy performance depends on the specific allocation of N_1, N_2 rather than on the total number of CPT symbols $N = N_1 + N_2$ alone, thus, motivating a proper optimization of the DL and UL duration. Specifically, we illustrate that short DL phases are preferable in highly sparse networks (with small realizations of K) given a fixed N. Moreover, we motivate the proposed CPT + MUD over the conventional standalone MUD implementation by presenting and discussing some preliminary results on their attainable MUD performance.

Finally, we discuss several attractive research directions related to CPT to pursue in the sequence.

C. Organization

The remainder of this paper is organized as follows. Section II presents the system model and introduces the proposed CPT mechanism and signal sparsity level estimator. The accuracy of the proposed estimator is assessed in Section III, while Section IV discusses numerical results. Finally, Section V concludes the article and highlights further research directions.

Notation: Boldface lowercase letters denote column vectors. Superscripts $(\cdot)^*$ and $(\cdot)^H$ denote the complex conjugate and Hermitian operations, respectively. $||\cdot||$ is the Euclidean norm of a vector, $|\cdot|$ is the absolute (or cardinality for sets) operation, and $\mathrm{round}(\cdot)$ denotes the rounding-to-the-nearest rounding operation. $\Pr[A]$ denotes the probability of the occurrence of event A, while $A|B$ denotes a random variable A conditioned on B. $\mathbb{E}[\,\cdot\,]$ and $\mathrm{var}[\,\cdot\,]$ output the expected value and variance of the argument, respectively. $\Re\{\cdot\}$ ($\Im\{\cdot\}$) outputs the real (imaginary) part of the argument. Additionally, $\mathrm{E}_1(\cdot)$ is the exponential integral [52, eq. (6.2.1)], $\mathrm{erfc}(\,\cdot\,)$ is the complementary error function [52, eq. (7.2.2)], and $\Gamma(\,\cdot\,)$ is the complete gamma function [52, eq. (5.2.1)]. $\mathbb{C}$ is the set of complex numbers, and $\imath = \sqrt{-1}$ is the imaginary unit. $f_X(x)$ and $F_X(x)$ denote the probability density function (PDF) and cumulative distribution function (CDF), respectively, of a continuous random variable X, while $p_Y(y)$ denotes the probability mass function (PMF) of a discrete random variable Y. Moreover, $X \sim \mathcal{C}\big(\mathbb{E}[X], \mathrm{var}[X]\big)$, $X \sim \mathcal{CN}\big(\mathbb{E}[X], \mathrm{var}[X]\big)$, $X \sim \mathrm{Ray}\big(\mathbb{E}[X]\sqrt{2/\pi}\big)$, and $X \sim \mathcal{T}(\nu)$, are respectively a Gaussian, a circularly-symmetric complex Gaussian, a Rayleigh, and a Student's t with ν degree of freedom, random variable. Finally, Table I lists the main symbols used throughout the paper.

[1]Notice that the estimation of the number of receive radio frequency signals has been already investigated in the literature for some decades (see [45]–[49]). However, the proposed approaches require a sufficiently large number of available samples/symbols, which is affordable for the considered radio-cognitive, spectrum sensing, and radar applications but not for mMTC, where messages are natively short. Therefore, they cannot be leveraged to accurately estimate the signal sparsity level in mMTC so as to be beneficial for MUD.

[2]By convention [50], [51], a detection or classification operation is applied over a (discrete) set of possible hypotheses, while an estimation operation is not restricted to a discrete/natural domain. Hence, a detector for K outputs an integer solution, while an estimator for K may output a real solution.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2023.3337242

TABLE I: Main symbols used throughout the paper

Symbol	Definition
$\mathcal{Q}$ (Q)	set (total number) of devices in the network
$\mathcal{K}$ (K)	set (number) of simultaneously active devices in a time slot
h_i	$i-$th channel coefficient ($i-$th device $\rightleftarrows$ coordinator)
$\hat{h}_i$ ($\tilde{h}_i$)	estimator (estimation error) of h_i
β_i	average channel power gain of the $i-$th channel
N	total number of CPT symbols for the purpose of detecting K
N_1 (N_2)	number of symbols for DL (UL) phases. $N_1 + N_2 = N$
$\mathbf{v}$ ($\mathbf{s}$)	DL (UL) pilot CPT signal
$\mathbf{z}_i$ ($\mathbf{y}$)	receive signal at the $i-$th device (coordinator) in the DL (UL)
w_i	additive white Gaussian noise (AWGN) in the $i-$th device
w	AWGN at the coordinator
σ^2	variance of the AWGN in the devices and coordinator
p (ρ)	transmit power (target receive power) of (at) the coordinator
$\varphi_{\mathcal{K}}$	uncertainty in the UL receive signal related to $\{\tilde{h}_i\}_{\forall i \in \mathcal{K}}$
$\widehat{K}$ ($\widehat{K}_r$)	integer (relaxed/real) estimator of K
ϱ	DL transmit signal-to-noise ratio (SNR)
λ_i (ϑ_i)	power (variance) of $\hat{h}_i$
μ	channel power transmission threshold in the devices
$p_{\max}$ ($\bar{p}_i$)	permissible maximum (average) device transmit power
$p_{\text{out},i}$	probability that an active device cannot transmit
$\bar{\gamma}_u$	target receive SNR in the UL
ϑ	smallest possible ϑ_i, i.e., $\vartheta \triangleq \min_{i \in \mathcal{Q}} \vartheta_i$
ζ	detection threshold in the AMP MUD mechanism

II. System Model & Proposed CPT

Consider an mMTC deployment, where a set $\mathcal{Q}$ of devices is served by a single coordinator, e.g., a base station or an aggregator. It is assumed that all devices and the coordinator are equipped with a single antenna.[3] Assume that time is slotted and active devices must wait for the next immediate time slot to start a (synchronous) transmission. Let us denote by h_i the channel coefficient of the link between the coordinator and the $i-$th device, and assume that the channels are subject to quasi-static fading and remain unchanged during each time slot. In addition, DL and UL channels are reciprocal, which is motivated by the use of the same frequency band and a TDD operation [4], [6], [39].

The MTC traffic is sporadic, i.e., only a random subset of the devices $\mathcal{K} \subseteq \mathcal{Q}$ is active at any given time. We aim to detect the number of devices $K = |\mathcal{K}|$, out of the total $Q = |\mathcal{Q}|$, becoming active in a given time slot, which is also referred to as signal sparsity level. This information is then potentially exploited in posterior detection/decoding mechanisms, e.g., [13]–[35], [37]–[44].

As illustrated in Fig. 1, the proposed CPT mechanism for estimating the sparsity level, K, consists of a DL and a UL pilot transmission phase. This is followed by the transmission of training and data symbols in the case of coherent MUD, or only data symbols in the case of non-coherent MUD.

A. DL Phase

At the beginning of each time slot, the coordinator sends a broadcast pilot signal $\mathbf{v} \in \mathbb{C}^{N_1}$ comprising N_1 symbols. The

[3]As this is, to the best of our knowledge, the first work that proposes the sparsity level detection problem prior to the MUD, our aim here lies in introducing the basic ideas, principles, and performance baselines, and we focus on single antenna devices. The extension of our proposed mechanisms to multi-antenna setups is not only an interesting but a required research direction, which demands specific but non-trivial adjustments and analyses that we leave for future work.

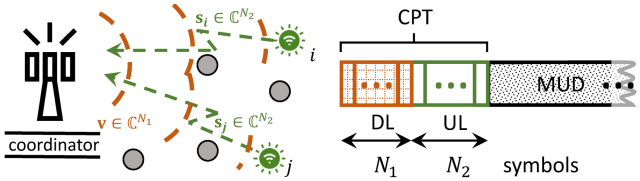

Fig. 1: Proposed CPT mechanism.

signal received by the $i-$th device is given by

$$z_i[n] = \sqrt{p} h_i v[n] + w_i[n], \; n = 1, 2, \cdots, N_1, \tag{1}$$

where $||\mathbf{v}||^2 = N_1$, p is the per-symbol average transmit power of the coordinator, and $w_i[n] \sim \mathcal{CN}(0, \sigma_i^2)$ is the $n-$th AWGN sample at the $i-$th device. For simplicity, we assume $\sigma_i^2 = \sigma^2, \forall i$.

This DL broadcast pilot transmission phase is leveraged by the active devices to estimate their corresponding channel coefficient since the UL and DL channels are reciprocal. Specifically, the minimum variance unbiased estimator of h_i and the corresponding estimation error are respectively given by

$$\hat{h}_i = \mathbf{v}^H \mathbf{z}_i / (N_1 \sqrt{p}), \tag{2}$$

$$\tilde{h}_i = \mathbf{v}^H \mathbf{w}_i / (N_1 \sqrt{p}), \tag{3}$$

with $\hat{h}_i = h_i + \tilde{h}_i$.

B. UL Phase

Note that the transmission of DL pilots for the acquisition of DL / UL CSI (and corresponding power control, precoding/beamforming design, and other channel-aware resource allocation mechanisms) is widely used in TDD systems [4], [6], [9].[4] The innovative part of our proposal lies in how this information is exploited for UL pilot transmissions. Specifically, we propose that, after the DL CSI acquisition phase, the active devices exploit the remaining N_2 symbols for sending a *common/shared* pilot sequence $\mathbf{s} \in \mathbb{C}^{N_2}$, with $|s[n]|^2 = 1 \; \forall n$, but phase shifted as $e^{-\imath \angle \hat{h}_i} \mathbf{s} = \hat{h}_i^* \mathbf{s} / |\hat{h}_i|$, thus, aiming at a coherent signal combination at the coordinator. We adopt a channel inversion power control such that the $i-$th device transmits with power $\frac{\rho}{|\hat{h}_i|^2}$ given a target receive power ρ. The signal received at the coordinator, $\mathbf{y} \in \mathbb{C}^{N_2}$, is given by

$$\begin{aligned} y[n] &= \sum_{i \in \mathcal{K}} \sqrt{\rho} \hat{h}_i^* h_i s[n] / |\hat{h}_i|^2 + w[n] \\ &= \sqrt{\rho} K s[n] + \varphi_{\mathcal{K}} s[n] + w[n], \end{aligned} \tag{4}$$

for $n = 1, 2, \cdots, N_2$, where $w[n] \sim \mathcal{CN}(0, \sigma^2)$ is the AWGN sample at the coordinator. Finally, the last step in (4) follows after using $\hat{h}_i = h_i + \tilde{h}_i$ and setting

$$\varphi_{\mathcal{K}} \triangleq \sqrt{\rho} \sum_{i \in \mathcal{K}} \frac{\hat{h}_i^* \tilde{h}_i}{|\hat{h}_i|^2}, \tag{5}$$

which denotes the uncertainty in the UL receive signal related to the CSI estimation error.

[4]In current fifth-generation wireless communications systems, reference and synchronization DL pilot signals, which are transmitted using a collection of broad and mildly-directional multi-antenna beams over several spectrum resource blocks, are used for this purpose [53].

C. Signal Sparsity Level Estimator

The signal $\mathbf{y}$ received at the coordinator is used to detect the signal sparsity level, K, among all the $Q+1$ possible hypotheses: $\mathcal{H}_k : K = k$, where $k = 0, 1, \cdots, Q$. Here, notice that the distribution and statistics of $\varphi_{\mathcal{K}}$ are completely unknown given no prior modeling assumption of the channel. Therefore, probably the wisest thing to do is to relax the integer detection problem to a real-domain continuous estimation as suggested by [51]. Moreover, observe that

$$\mathbb{E}[\mathbf{s}^H\mathbf{y}] = \sqrt{\rho}N_2K$$

since $\mathbb{E}[\tilde{h}_i] = 0$, thus $\mathbb{E}[\varphi_{\mathcal{K}}] = 0$. More specifically, $\mathbb{E}[\Re\{\mathbf{s}^H\mathbf{y}\}] = \sqrt{\rho}N_2K$, while $\mathbb{E}[\Im\{\mathbf{s}^H\mathbf{y}\}] = 0$. Thus, we can use the method of moments in estimation theory [50] to obtain the relaxed estimator

$$\widehat{K}_r = \Re\{\mathbf{s}^H\mathbf{y}\}/(N_2\sqrt{\rho}), \tag{6}$$

and set $\widehat{K} = \text{round}(\widehat{K}_r)$. Note that $\mathbb{E}[\widehat{K}] = K$, i.e., the estimator is unbiased. Good detection accuracy is expected if $\mathbb{E}[|\varphi_{\mathcal{K}}|^2], \sigma^2 \ll \rho K$, as should occur in practice by design.

Finally, note that our CPT proposal cannot be applied in setups where channel reciprocity does not hold, such as frequency division duplex systems.

III. Accuracy of the Proposed Estimator

In the following, we adopt a channel model for the purpose of assessing the detection accuracy of the proposed CPT mechanism and corresponding estimator. Specifically, channels are assumed to be subject to quasi-static Rayleigh fading such that $h_i \sim \mathcal{CN}(0, \beta_i)$, where β_i is the average channel power gain. Using this, together with (2) and (3), one obtains

$$\hat{h}_i \sim \mathcal{CN}(0, \vartheta_i), \tag{7}$$

$$\tilde{h}_i \sim \mathcal{CN}\Big(0, \frac{1}{N_1\varrho}\Big), \tag{8}$$

where $\varrho \triangleq p/\sigma^2$ is the DL transmit SNR, and $\vartheta_i \triangleq \beta_i + \frac{1}{N_1\varrho}$.

Moreover, we consider that the transmitting (thus, detectable) devices are only the active devices whose channels are not deeply faded, i.e., those with $\lambda_i \triangleq |\hat{h}_i|^2 \geq \mu$. In practice, μ must be set based on the permissible maximum power and/or average power consumption, which are given respectively by

$$p_{\max} = \rho/\mu, \tag{9}$$

$$\begin{aligned}\bar{p}_i &= \mathbb{E}\big[\rho/\lambda_i \;\big|\; \lambda_i \geq \mu\big] \\ &= \frac{\rho}{1 - F_{\Lambda_i}(\mu)} \int_{\mu}^{\infty} \frac{1}{\lambda_i} f_{\Lambda_i}(\lambda_i) \mathrm{d}\lambda_i \\ &= \frac{\rho}{\vartheta_i} \Gamma\Big(0, \frac{\mu}{\vartheta_i}\Big) e^{\frac{\mu}{\vartheta_i}},\end{aligned} \tag{10}$$

where the last step comes from exploiting that λ_i is an exponential random variable with mean ϑ_i, and using the definition of the upper incomplete gamma function [52, eq. 8.2.2]. Finally, the probability that an active device cannot transmit due to insufficient power to compensate for the channel losses is given by

$$p_{\text{out},i} = \Pr[\lambda_i < \mu] = F_{\Lambda_i}(\mu) = 1 - e^{-\mu/\vartheta_i}. \tag{11}$$

A. Variance of the Relaxed Estimator

In the following, we analyze the variance of the relaxed estimator (6). First, let us define $V \triangleq \Re\{\mathbf{s}^H\mathbf{w}/(N_2\sqrt{\rho})\}$ and depart from (6) to obtain

$$\begin{aligned}\text{var}[\widehat{K}_r] &= \text{var}\Big[\Re\Big\{\frac{\varphi_{\mathcal{K}}}{\sqrt{\rho}}\Big\}\Big||\hat{h}_i| \geq \sqrt{\mu}\Big] + \text{var}[V] \\ &\overset{(a)}{=} \left(\sum_{i\in\mathcal{K}} \text{var}\left[\Re\Big\{\frac{\hat{h}_i^*\tilde{h}_i}{|\hat{h}_i|^2}\Big\}\Big||\hat{h}_i| \geq \sqrt{\mu}\right] + \frac{1}{2N_2\bar{\gamma}_u}\right) \\ &\overset{(b)}{=} \sum_{i\in\mathcal{K}} \text{var}\left[\frac{\Re\{\tilde{h}_i\}}{|\hat{h}_i|}\Big||\hat{h}_i| \geq \sqrt{\mu}\right] + \frac{1}{2N_2\bar{\gamma}_u} \\ &\overset{(c)}{=} \sum_{i\in\mathcal{K}} \frac{\text{var}[Z_i]}{2N_1\varrho\vartheta_i} + \frac{1}{2N_2\bar{\gamma}_u},\end{aligned} \tag{12}$$

where (a) follows from using (5), setting $\bar{\gamma}_u \triangleq \rho/\sigma^2$, which denotes the target receive SNR in the UL, and using $\text{var}[V] = 1/(2N_1\bar{\gamma}_u)$ since $V \sim \mathcal{N}(0, 1/(2N_2\bar{\gamma}_u))$. Meanwhile, (b) comes after exploiting the fact that $\hat{h}_i^*\tilde{h}_i/|\hat{h}_i|^2$ is equivalently distributed as $\tilde{h}_i/|\hat{h}_i|$ since $\hat{h}_i^*/|\hat{h}_i|$ is uniformly distributed in the unit circle and independent of $\tilde{h}_i/|\hat{h}_i|$, thus, it does not alter the latter's distribution. Finally, (c) uses $Z_i \triangleq \sqrt{2N_1\varrho\vartheta_i}\Re\{\tilde{h}_i\}/|\hat{h}_i| \;\big|\; |\hat{h}_i| \geq \sqrt{\mu}$.

Observe that Z_i can be written as $Z_i = X_i/Y_i$, where $X_i \triangleq \sqrt{2N_1\varrho}\Re\{\tilde{h}_i\} \sim \mathcal{N}(0,1)$, and $Y_i \triangleq |\hat{h}_i|/\sqrt{\vartheta_i} \;\big|\; |\hat{h}_i| \geq \sqrt{\mu}$. Therefore, by letting $U \triangleq |\hat{h}_i|/\sqrt{\vartheta_i} \sim \text{Ray}(1/\sqrt{2})$ and noticing that $Y_i \sim U \mid U \geq \sqrt{\mu/\vartheta_i}$, one obtains

$$\begin{aligned}f_{Y_i}(y) &= \frac{f_U(y)}{1 - F_U(\sqrt{\mu/\vartheta_i})} \\ &= 2ye^{\mu/\vartheta_i - y^2} \quad \text{for } y \geq \sqrt{\mu/\vartheta_i}.\end{aligned} \tag{13}$$

Then, since X_i and Y_i are independent and $\mathbb{E}[X_i] = 0$, we proceed as follows

$$\begin{aligned}\text{var}[Z_i] &= \mathbb{E}[X_i^2]\mathbb{E}[Y_i^{-2}] \\ &\overset{(a)}{=} \int_0^{\infty} y^{-2} f_{Y_i}(y) \mathrm{d}y \\ &\overset{(b)}{=} \int_{\sqrt{\mu/\vartheta_i}}^{\infty} \frac{2}{y} e^{\mu/\vartheta_i - y^2} \mathrm{d}y \\ &\overset{(c)}{=} e^{\frac{\mu}{\vartheta_i}} \mathrm{E}_1(\mu/\vartheta_i),\end{aligned} \tag{14}$$

where (a) comes from using $\mathbb{E}[X_i^2] = 1$ and the integral form of $\mathbb{E}[Y_i^{-2}]$, (b) comes from substituting (13), while (c) is attained by applying simple algebraic transformations and using the definition of the exponential integral [52, eq. (6.2.1)].

Now, by substituting (14) into (12), one attains

$$\text{var}[\widehat{K}_r] = \frac{1}{2N_1\varrho\mu} \sum_{i\in\mathcal{K}} g(\mu/\vartheta_i) + \frac{1}{2N_2\bar{\gamma}_u}, \tag{15a}$$

$$\leq \frac{g(\mu/\vartheta)}{2N_1\varrho\mu} K + \frac{1}{2N_2\bar{\gamma}_u}, \tag{15b}$$

where $g(x) = xe^x\mathrm{E}_1(x)$. Meanwhile, in the last line, we use $\vartheta \triangleq \min_{i\in\mathcal{Q}} \vartheta_i$ motivated by the fact that $\text{var}[\widehat{K}_r]$ is a decreasing function of ϑ_i. This can be corroborated by noticing

that g is an increasing function of x as both bounds in [52, eq.6.8.1]

$$\frac{x}{2}\ln(1+2/x) < g(x) < x\ln(1+1/x) \tag{16}$$

increase with x. All this shows that the worst-case scenario is where the active devices are the farthest from the coordinator and, thus, are characterized by the smallest ϑ_i.

Hereinafter, we focus on the worst-case deployment scenario, i.e., the active devices are at the edge of the service area such that $\vartheta_i = \vartheta,\ \forall i \in \mathcal{K}$. Then, using (15) and (16), one obtains

$$\mathrm{var}[\widehat{K}_r] = \frac{g(\mu/\vartheta)}{2N_1\varrho\mu}K + \frac{1}{2N_2\bar{\gamma}_u}, \tag{17a}$$

$$\mathrm{var}[\widehat{K}_r] < \frac{\ln(1+\vartheta/\mu)}{2N_1\varrho\vartheta}K + \frac{1}{2N_2\bar{\gamma}_u}, \tag{17b}$$

$$\mathrm{var}[\widehat{K}_r] > \frac{\ln(1+2\vartheta/\mu)}{4N_1\varrho\vartheta}K + \frac{1}{2N_2\bar{\gamma}_u}. \tag{17c}$$

Finally, $\mathrm{var}[\widehat{K}] > \mathrm{var}[\widehat{K}_r]$ due to the variance introduced by the rounding operation. Interestingly, such additional variance is always smaller than $1/12$ as this corresponds to the worst-case scenario, where $\widehat{K} - \widehat{K}_r$ is uniformly distributed in $[-1/2, 1/2]$.

B. Distribution of the Estimator

In general, and especially for the considered setup, the estimator variance cannot be directly used to quantify, at least thoroughly, the performance degradation due to detection mismatches. Instead, the distribution of the classification results must be taken into account.

The PMF of $\widehat{K}$ can be found as

$$\begin{aligned} p_{\widehat{K}}(\hat{k}) &= \Pr\left[\mathrm{round}(\widehat{K}_r) = \hat{k}\right] \\ &= \Pr\left[\hat{k} - 1/2 \le \widehat{K}_r \le \hat{k} + 1/2\right] \\ &= F_{\widehat{K}_r}(\hat{k}+1/2) - F_{\widehat{K}_r}(\hat{k}-1/2). \end{aligned} \tag{18}$$

Notice that the distribution of the relaxed estimator, $\hat{K}_r$, is needed for computing $p_{\widehat{K}}(\hat{k})$. Hence, the problem translates to finding $F_{\widehat{K}_r}(\hat{k})$, for which we rewrite (6) as

$$\widehat{K}_r = K + \sum_{i\in\mathcal{K}} \frac{Z_i}{\sqrt{2N_1\varrho\vartheta}} + V, \tag{19}$$

and proceed as follows.

The distribution of Z_i is derived as

$$\begin{aligned} f_{Z_i}(z) &= \frac{d}{dz}\Pr\left[\frac{X_i}{Y_i} \le z\right] = \frac{d}{dz}\int_{\sqrt{\mu/\vartheta}}^{\infty} F_{X_i}(yz) f_{Y_i}(y)\mathrm{d}y \\ &\overset{(a)}{=} \int_{\sqrt{\mu/\vartheta}}^{\infty} y f_{X_i}(yz) f_{Y_i}(y)\mathrm{d}y \\ &\overset{(b)}{=} \sqrt{\frac{2}{\pi}}e^{\mu/\vartheta}\int_{\sqrt{\mu/\vartheta}}^{\infty} y^2 e^{-y^2(z^2/2+1)}\mathrm{d}y \\ &\overset{(c)}{=} \frac{e^{\mu/\vartheta}}{\sqrt{2\pi}}\left(\frac{z^2}{2}+1\right)^{-3/2}\Gamma\left(\frac{3}{2}, \left(\frac{z^2}{2}+1\right)y^2\right)\Big|_{y=\sqrt{\mu/\vartheta}}^{y\to\infty} \\ &\overset{(d)}{=} \frac{e^{\mu/\vartheta}\Gamma\left(3/2, \mu(z^2/2+1)/\vartheta\right)}{\sqrt{2\pi(z^2/2+1)^3}}, \end{aligned} \tag{20}$$

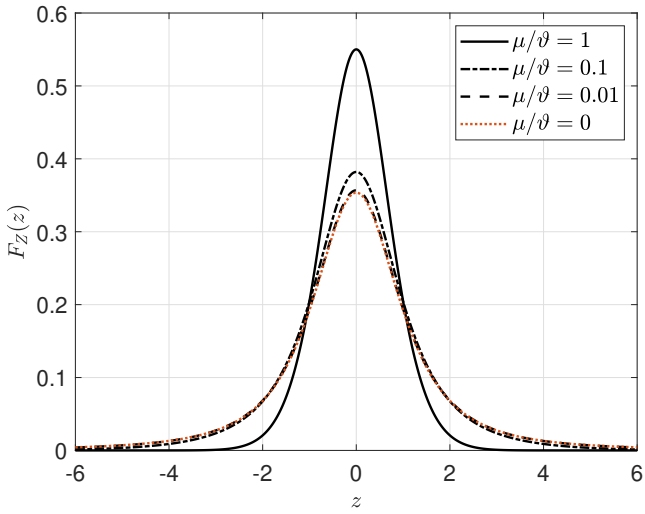

Fig. 2: PDF of Z according to (20) for $\mu/\vartheta \in \{0, 0.01, 0.1, 1\}$.

where (a) comes from differentiating under the integral sign by leveraging Leibniz rule and $dF_{X_i}(yz)/dz = yf_{X_i}(yz)$, and (b) follows from substituting $f_{X_i}(x) = e^{-x^2/2}/\sqrt{2\pi}$, and $f_{Y_i}(y)$ as given in (13). The indefinite integral is solved in (c) using [54, eq. (2.325.6)], while (d) follows directly after evaluating the definite integral limits. Fig. 2 illustrates the shape of $f_{Z_i}(z)$ for different values of μ/ϑ. By using (11), one obtains $p_{\text{out}} = 0.6321$ already for $\mu/\vartheta = 1$, thus, we only considered configurations with $\mu/\vartheta \le 1$, which are required to guarantee a relatively small $p_{\text{out},i}$. Notice that $f_Z(z)$ is symmetric around 0, which is expected since X_i is a zero-mean Gaussian random variable and $Y_i \ge 0$, and is bell-shaped.

With (20) at hand, the CDF of $\widehat{K}_r$ can be obtained as follows

$$\begin{aligned} F_{\widehat{K}_r}(\hat{k}) &= \Pr\left[K + \frac{1}{\sqrt{2N_1\varrho\vartheta}}\sum_{i\in\mathcal{K}} Z_i + V \le \hat{k}\right] \\ &= \underbrace{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}}_{K \text{ integrals}} F_V\left(\hat{k} - K - \sum_{i\in\mathcal{K}}\frac{z_i/\sqrt{2}}{\sqrt{N_1\varrho\vartheta}}\right)\prod_{i\in\mathcal{K}} f_{Z_i}(z_i)\mathrm{d}z_i, \end{aligned} \tag{21}$$

where $F_V(v) = 1 - \mathrm{erfc}(v\sqrt{N_2\bar{\gamma}_u})/2$. Unfortunately, evaluating (21) becomes computationally expensive and often unaffordable, especially when $K \gg 1$ due to the increased number of integration operations.

To address the above issue, herein we exploit the fact that $T \triangleq \sum_{i\in\mathcal{K}} Z_i$ is approximately distributed as a scaled Student's t distribution $\sqrt{w_1(1-2/\nu)}\mathcal{T}(\nu)$, where ν is the solution to

$$2^{\frac{\nu}{2}-1}\omega_2\Gamma\left(\frac{\nu}{2}\right) = K_{\frac{\nu}{2}}\left(\sqrt{\omega_1(\nu-2)}\right)\left(\sqrt{\omega_1(\nu-2)}\right)^{\frac{\nu}{2}}, \tag{22}$$

and

$$\omega_1 \triangleq e^{\mu/\vartheta}\mathrm{E}_1(\mu/\vartheta)K, \tag{23}$$

$$\omega_2 \triangleq \left[2\int_0^{\infty}\cos(z) f_Z(z)\mathrm{d}z\right]^K. \tag{24}$$

See the Appendix for the proof and accuracy-related discussions. From (19), this implies that the distribution of

the relaxed estimator is symmetric around K and accurately matches the distribution of the sum of a Gaussian and a Student's t random variable.

Now, let us denote $T' \sim \mathcal{T}(\nu)$, where ν is the solution to (22). Then, one attains

$$\begin{aligned} F_{\widehat{K}_r}(\hat{k}) &\overset{(a)}{\approx} \Pr\left[V \leq \hat{k}-K-\sqrt{\frac{\omega_1(1-2/\nu)}{2N_1\varrho\vartheta}}T'\right] \\ &= \int_{-\infty}^{\infty} F_V\left(\hat{k}-K-\sqrt{\frac{\omega_1\left(1-\frac{2}{\nu}\right)}{2N_1\varrho\vartheta}}t\right) f_T(t)\mathrm{d}t \\ &\overset{(b)}{=} 1-\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{2\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\int_{-\infty}^{\infty}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \\ &\quad \times \operatorname{erfc}\left(\sqrt{N_2\bar{\gamma}_u}\left(\hat{k}-K-\sqrt{\frac{\omega_1\left(1-\frac{2}{\nu}\right)}{2N_1\varrho\vartheta}}t\right)\right)\mathrm{d}t, \quad (25) \end{aligned}$$

where (a) comes from using the first line of (21) together with the definition of T, while (b) follows from using the distribution of V and T'. Notice that by using [55, eq. (3)], one can state the integral operation in (25) as an infinite sum that includes factorials, incomplete gamma, and confluent hypergeometric functions. However, such an approach may not significantly reduce the mathematical complexity of numerical computing (25), so we do not adopt it here.

Finally, observe that computing (25) is much less computationally demanding than (21) since only two integrals must be evaluated, i.e., (25) and ω_2 in (24), independently of the value of K.

IV. Numerical Results

In this section, we numerically analyze the performance of the proposed CPT mechanism under Rayleigh fading channel conditions. For this, we resort to Monte Carlo simulations and the analytical approximations derived in Section III, which are shown to match closely. Performance is evaluated in terms of the detection success probability given by $p_{\widehat{K}}(K)$, which can be approximately obtained from evaluating (18) using (25), and the variance of the relaxed estimator, which is analytically characterized in (17).

We consider the worst-case deployment scenario, where all devices are at the edge of the service area, so $\beta_i = \beta\ \forall i$. Unless stated otherwise, we consider a massive deployment of $Q = 1000$ devices, out of which $K = 5$ become active at each time slot, and $\beta = -120$ dB, which may correspond to a link distance in the order of 500 m [29]. Also, we set $N_1 = 2$ and $N_2 = 4$ such that $N = 6$ symbols are dedicated to CPT. This may be a reasonable choice considering that the overall transmission time may comprise many more symbols depending on the data traffic and connectivity solution.[5] Let $p = 30$ dBm, $\mu = -140$ dB, and $\rho = -115$ dBm such that the maximum power allowed (9) and the average power consumption (10) of the devices are 316.2 mW and 12.9 mW, respectively, while the probability that an active device cannot transmit due to insufficient power to compensate for the channel losses is 10^{-2}. Finally, we set $\sigma^2 = -120$ dBm by assuming a transmission bandwidth of 180 kHz.

[5]For instance, orthogonal frequency division multiplexing-based solutions such as NB-IoT and LTE-M may support 14 symbols per subcarrier in a 1.17 ms time slot of 15 KHz bandwidth [56], [57]. Message transmissions may span over several time slots.

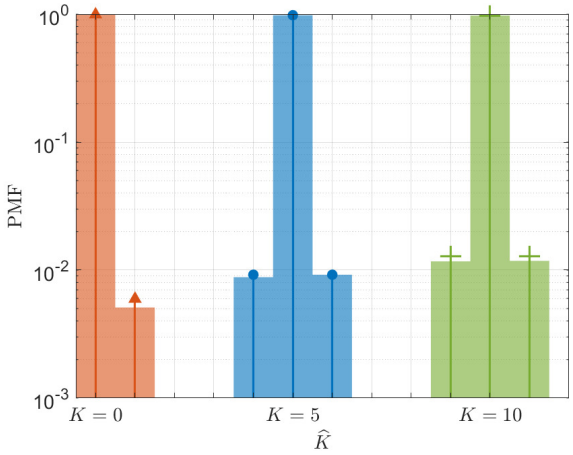

Fig. 3: PMF of $\widehat{K}$ for $K \in \{0, 5, 10\}$. Markers correspond to the analytical PMF approximation.

A. On the Detection Scalability

Fig. 3 depicts the PMF of $\widehat{K}$ considering several values of K. The results here corroborate the symmetric shape of the distribution of $\widehat{K}_r$ (and $\widehat{K}$), which matches approximately that of the sum of a Student's t and a Gaussian random variable. Moreover, the accuracy of the estimation decreases with K. The latter phenomenon can be more clearly appreciated in Fig. 4, where both the variance of the relaxed estimator (Fig. 4a) and the corresponding detection success probability (Fig. 4b) are plotted against K for $\beta \in \{-130, -120\}$ dB. Indeed, the variance of the relaxed estimator increases linearly with K and decreases with β as predicted by (17), while being lower-bounded by the noise variance level, i.e., $\frac{1}{2N_1\bar{\gamma}_u}$. The quantitative impact of such behavior is captured by the detection success probability metric, which shows, for instance, that the signal sparsity level, K, is predicted with an accuracy of 91%, and 98% for $K = 5$ when $\beta = -130$ dB, and $\beta = -120$ dB, respectively, while such figures decrease to 76%, and 97% when $K = 15$.

B. How Many CPT Symbols are Needed?

Fig. 5 shows the detection success probability as a function of N_1 for a fixed number of CPT symbols $N = 6$. This is, $N_1 + N_2 = N$, thus, $N_2 = N - N_1$. Note that the allocation of the DL/UL symbol significantly influences the detection success probability. Indeed, a relatively small/large ρ makes the DL phase less/more performance sensitive, thus motivating the allocation of less/more pilots to it for optimum performance. For instance, the optimum pilot allocation is $N_1 = 1$ and $N_1 = 3$ when $\rho = -115$ dBm and $\rho = -110$ dBm, respectively. Observe that the optimum configuration of N_1, N_2 is also the one that minimizes $\mathrm{var}[\widehat{K}_r]$ since the estimator's distribution is symmetric around K. Since N_1, N_2 are

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2023.3337242

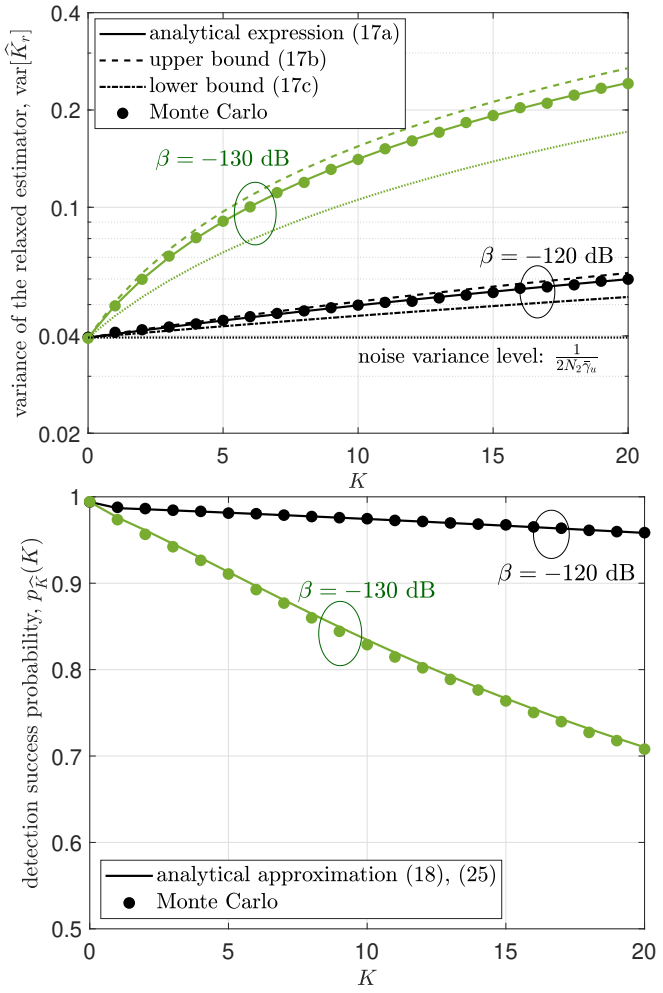

Fig. 4: a) Variance of the relaxed estimator (6) (top), and b) detection success probability (bottom), as a function of K. We set $\mu = \beta/100$ such that $p_{\text{out}} \approx 10^{-2}$ independently of β.

positive integers and N is usually small in practical setups, a brute force mechanism suffices to solve $\arg\min_{N_1,N_2} \text{var}[\widehat{K}_r]$ subject to $N_1 + N_2 = N$.

Meanwhile, Fig. 6a illustrates the potential performance improvements from increasing the total number of CPT symbols N. Herein, we test the performance under all possible combinations of (N_1, N_2) with $N_1 + N_2 = N$ and select the one leading to the best detection success probability (or simply, minimum estimator's variance), which appears depicted in Fig. 6b. Observe that the probability of detection success increases rapidly with N, which represents the degrees of freedom to resolve the uncertainties related to fading in the system. Moreover, as K increases, it is more beneficial to allocate more symbols to the DL phase, which is a behavior that can be deduced from (17). Specifically, the first term of (17) increases (decreases) with K (N_1), thus, these values must be traded-off for best performance. However, the value of K is not known beforehand; therefore, in practice, the optimization of N_1, N_2 can only be performed based on the statistical expectations of the number of active users K.

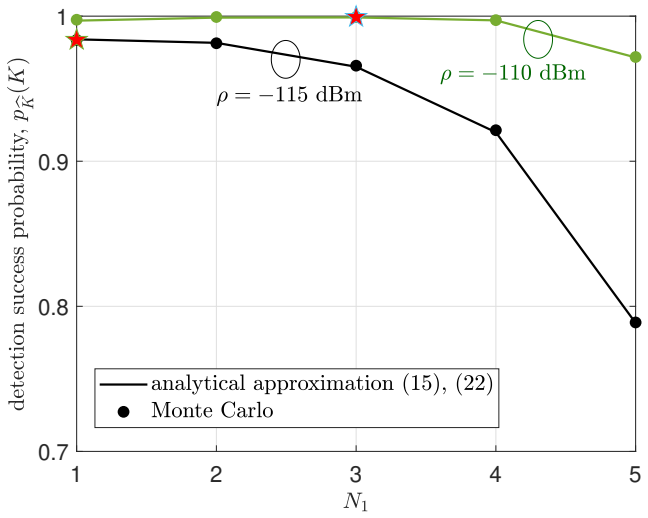

Fig. 5: Detection success probability as a function of N_1. We set $N_1 + N_2 = N$, thus, $N_2 = N - N_1$, with $N = 6$.

C. A Primer on the Performance of CPT + MUD

Next, we briefly illustrate how the performance of MUD, which comprises only UAD for simplicity, would benefit from incorporating the CPT mechanism proposed in this paper. For this, we consider that the coordinator is equipped with a 64-antenna array, although a single antenna is used for the purpose of signal sparsity level estimation using CPT. We adopt two fundamental CS-MUD algorithms: OMP [21], [23] and AMP [29], [31].[6] In the CPT + MUD implementation, K is first estimated using N symbols, then, MUD is executed employing M symbols in such a way that only the $\widehat{K}$ devices with the strongest estimated channel power, in the case of AMP, or the $\widehat{K}$ devices first detected, in the case of OMP, are declared active. We compare our proposed approach with the standard standalone MUD implementations leveraging $N+M$ symbols, where a device is detected if its associated estimated channel power, in the case of AMP, or the residual signal power immediately before detecting the device, in the case of OMP, exceeds a pre-defined threshold ζ. In the following, we assume that the devices use Bernoulli pilots as in [29] and become active with probability $\epsilon = 0.01$, thus, there are $\epsilon Q = 10$ devices active on average in the network.

Fig. 7 shows the activity detection error rate as a function of the detection threshold ζ, which is only used in the standalone MUD implementations. Here, a relatively small ζ tends to decrease the miss-detection probability but at the expense of more false-alarm events. In comparison, a relatively large ζ tends to decrease the false-alarm probability at the expense of the occurrence of more miss-detection events. Indeed, notice that as $\zeta \to 0$ and $\zeta \to \infty$, the activity detection error

[6]Note that there are more advanced MUD algorithms in the literature, some of which were briefly discussed in Section I. Herein, we focus on basic MUD algorithms for baseline integration and comparison, which allows us to establish a strong foundation for arguing the effectiveness or not of our proposal. In fact, basic MUD algorithms like OMP and AMP are well-understood and transparent, making it easier to reproduce, analyze, and interpret the results. The AMP implementation adopted here leverages a minimum mean square error-based (Bayes optimal) denoiser.

This article has been accepted for publication in IEEE Transactions on Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2023.3337242

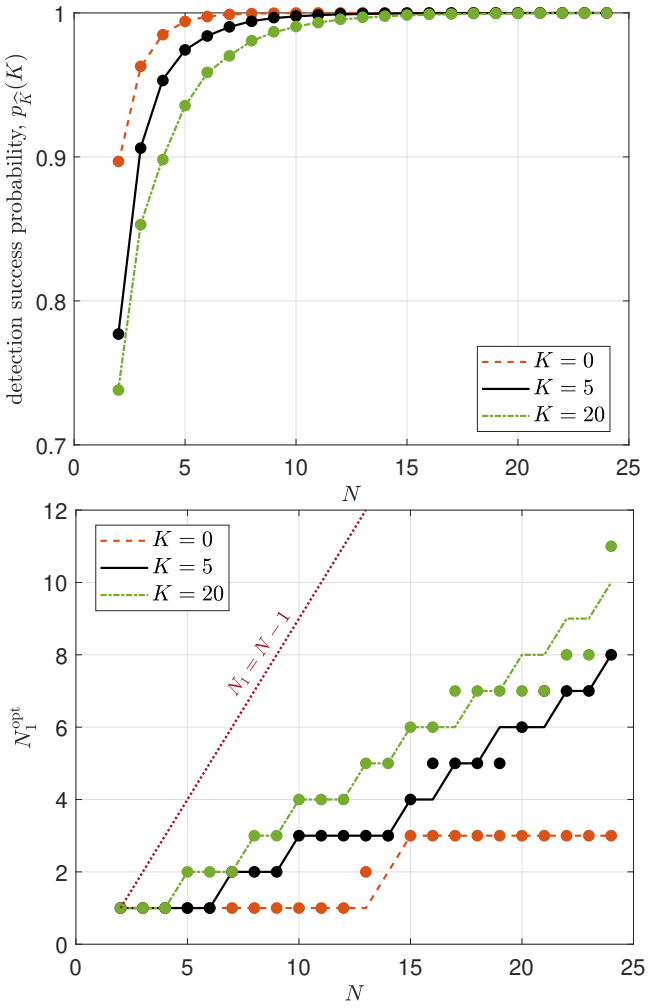

Fig. 6: a) Optimum detection success probability (top), and b) optimum N_1 as a function of N, thus, $N_2^{\text{opt}} = N - N_1^{\text{opt}}$. Here, the markers denote the results from Monte Carlo simulations.

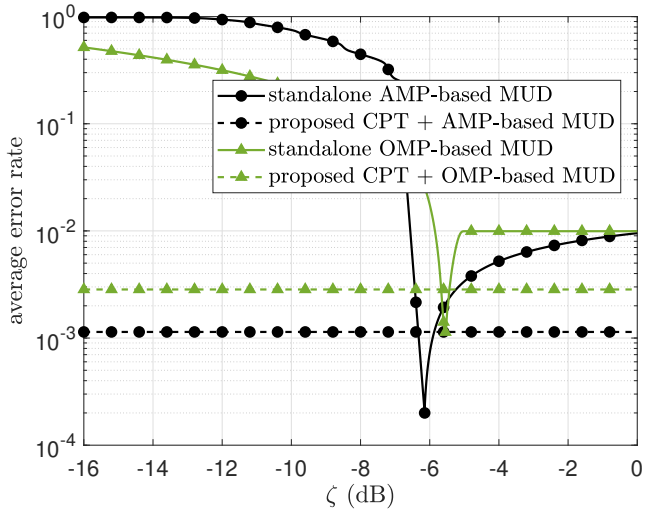

Fig. 7: Average activity detection error rate as a function of the detection threshold ζ. We set $N = 6$ and $M = 18$.

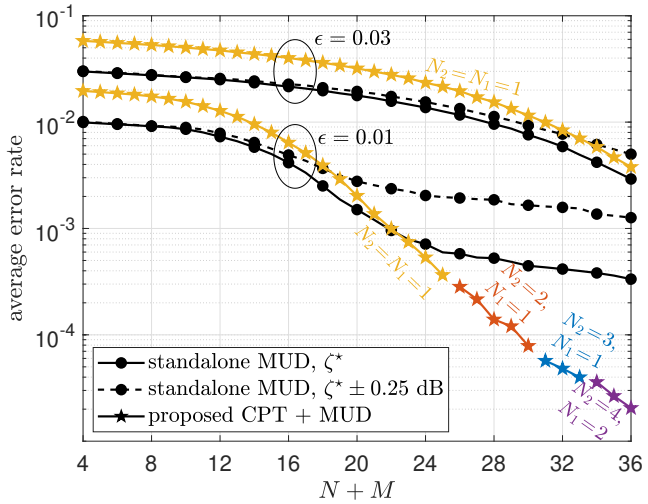

Fig. 8: Average activity OMP-based detection error rate as a function of the total number of symbols $N + M$. The results correspond to the optimum configuration of N_1, N_2, and M.

converges to $1 - \mathbb{E}[K]/Q = 0.99$ and $\mathbb{E}[K]/Q = 0.01$, respectively. This motivates the need to carefully tune ζ for optimum performance as discussed in Section I-A. Meanwhile, the proposed CPT + MUD completely avoids this tuning problem and, according to Fig. 7, can significantly outperform the standalone MUD implementations if the latter are not optimally/ideally configured as in practice. Note that the AMP outperforms OMP-based in terms of average error rate for relatively appropriate threshold choices, although at the cost of greater complexity and convergence time.

Fig. 8 illustrates the performance of standalone and CPT-assisted MUD mechanisms as a function of the number of symbols $N + M$. In this case, we adopt only OMP-based MUD algorithms for simplicity and illustrate the results corresponding to the optimum selection of N_1, N_2, and M. In the case of the standalone MUD mechanisms, we consider two configurations: one where the detection threshold ζ is optimal, i.e., $\zeta = \zeta^\star$, which corresponds to an ideal standalone configuration, and another where ζ can randomly deviate up to 0.25 dB from the optimal, i.e., ζ (dB)$\in$ [$\zeta^\star$ (dB)-0.25 dB, $\zeta^\star$ (dB)$+0.25$ dB]. As one may expect, when the number of available symbols is relatively small, the application of CPT is not advisable, and one may resort to standard standalone MUD approaches. However, as the number of available symbols increases, CPT assistance becomes more appealing. Interestingly, this depends on the sparsity of the network such that the number of symbols dedicated to CPT should be greater (smaller) for a smaller (greater) ϵ. Indeed, considering $\epsilon = 0.01$ and comparing with the ideal standalone OMP-based MUD, 0, 2, 3, 4, and 6 CPT symbols are recommended when the total number of symbols is [1,22], [23,25], [26,30], [31,33], [34,36], respectively. Meanwhile, when the system sparsity degrades such that $\epsilon = 0.03$, the MUD phase must be prioritized and CPT symbols may only be needed when the total number of available channels is as large as 33 considering a sub-optimal selection of ζ for a standalone MUD. Meanwhile, it always holds that $N_2 \geq N_1$ is preferable as also illustrated in Fig. 6. All in all, the results here evince that a CPT-assisted

MUD may provide significant performance gains relative to a standalone implementation, even if the detection threshold for the latter is optimally selected, considering the availability of a relatively large number of detection symbols.

V. Conclusion and Future Research Directions

In this work, we introduced a framework for detecting the number K of devices that become active, i.e., signal sparsity level, in an mMTC network. Specifically, the proposed CPT mechanism consists of a DL transmission using N_1 symbols for the purpose of CSI estimation, and an UL transmission with channel inversion (power and phase) control using N_2 shared symbols to resolve the fading uncertainty at the coordinator. We presented an efficient estimator for K based on such a UL signal and illustrated with some early results its crucial role for sparse signal recovery algorithms aiming at accurately identifying the specific set of active devices.

Regarding the signal sparsity level estimator, we analytically characterized its variance and distribution when the network channels are subject to Rayleigh fading. We showed that the estimator's variance increases linearly with K and that its distribution approximates that of the sum of a Student's t and a Gaussian random variable. The provided analytical framework allows tractable computation and optimization of the detection success probability, thus, becoming valuable for system design/analysis purposes. The numerical results showed that the attainable accuracy performance depends on the specific allocation of N_1, N_2 rather than on the total number of CPT symbols $N = N_1 + N_2$ alone, thus, motivating a proper optimization of the DL and UL phases. Indeed, we revealed that relatively short DL phases are preferable in highly sparse networks (with small realizations of K) given a fixed N.

To conclude, below we enumerate some attractive research directions that we would like to pursue in the sequence. Note that they aim to address key limitations of our current work such as the fact that the proposed CPT and signal sparsity level estimator are i) completely agnostic of the statistics of K, ii) designed only for single-antenna systems, iii) derived assuming perfect network synchronization, and iv) not jointly optimized with MUD.

1) Exploiting Prior Statistical Knowledge of K: We have not assumed any statistical knowledge of K. In practice, the coordinator might have some prior expectations based on traffic history, which can be leveraged for more accurate CPT-based detectors. In future work, our aim is to design CPT-based detectors exploiting traffic history.

2) CPT Optimized for MIMO Systems: MIMO technology is key for successful MUD, especially in mMTC networks with sporadic device activations. Therefore, adapting our proposed CPT framework to MIMO setups is undoubtedly appealing. Due to the overhead introduced by multi-antenna CSI training and the limited number of CPT symbols that may be available, an efficient proposal can rely on a compressed CPT training phase that limits the number of communicating antennas and/or exploits efficiently configured precoders/combiners.

3) CPT for Imperfectly Synchronized Networks: The proposed CPT mechanism requires network synchronization, as assumed throughout the paper. However, IoT devices may lack precise clocks and have heterogeneous hardware capabilities and protocol stacks, making it challenging to achieve accurate network synchronization [1], [2]. Mitigating the impact of timing discrepancies, especially for critical IoT networks, e.g., supporting URLLC [58], typically involves implementing time synchronization protocols and error-handling mechanisms, but perfect synchronization may not be achieved. Therefore, it is interesting to investigate/analyze the performance of CPT under imperfect synchronization conditions and propose synchronization error countermeasures, if applicable.

4) Joint CPT & MUD Optimization: The proposed CPT mechanism spanning over N symbols and aiming to determine the number of active devices is followed by MUD occupying M symbols, where the specific set of active devices is detected. Note that the number of active devices detected by CPT works as a prior for MUD mechanisms. An interesting question that we aim to address in future work is how to efficiently allocate the CPT and MUD symbols given that $N + M$ is constrained. For this, one needs to jointly assess the performance of both CPT and MUD mechanisms, which ultimately reveals the (practical) achievable performance of MUD. Some early insights were provided in the discussions around Fig. 8, but dedicated research and trade-off analysis are still required. Finally, comparisons with state-of-the-art MUD approaches that intrinsically implement signal sparsity level estimation, e.g., [26], [27], must be conducted.

Appendix

Recall that each Z_i is symmetric around 0 and bell-shaped as shown in Fig. 2 and discussed after (20). Therefore, T is also symmetric around 0 and bell-shaped. This suggests that a Student's t distribution, which is more general than a Gaussian distribution, may be a good fit. Several simulation campaigns that we carried out revealed that this is indeed the case.

At least two moments of T are needed to match those of a scaled Student's t distribution since such distribution is characterized only by the scale s and the number of degrees of freedom ν. The challenge lies in that ν must be greater than the moment order, while odd moments cannot be used since they are 0. For instance, this implies that we cannot rely on the first moment, and we cannot fit moments of order higher than 2 in order to allow $\nu \in (2, \infty)$ (which is required for having defined variance as is the case here, see (17)). The latter issue is very important since simulation results evinced that T may accurately fit, in many cases, a scaled Student's t distribution with ν approaching 2 from above. To avoid these issues, we resort to a fitting based on the second moment and characteristic function.

The second moment of T is given by

$$\omega_1 \triangleq \mathbb{E}[T^2] = K\mathbb{E}[Z^2], \tag{26}$$

which equals ω_1 in (24) by exploiting the independence and zero-mean features of $\{Z_i\}$ together with (14) with $\vartheta_i = \vartheta$

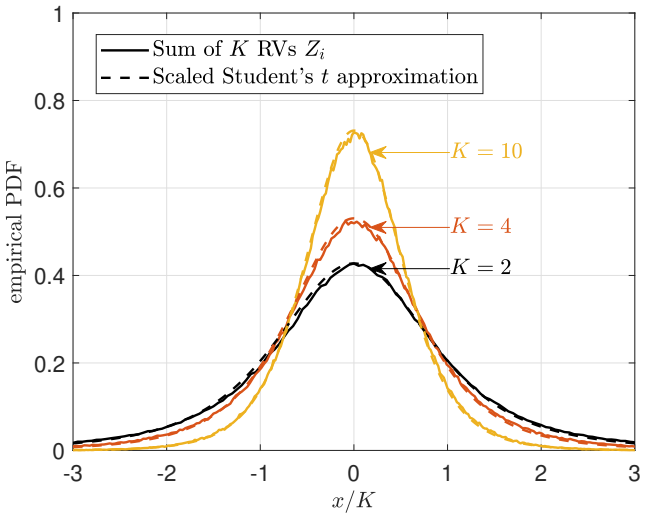

Fig. 9: Empirical PDF of the normalized sum of K i.i.d. RVs distributed as (20) and corresponding scaled Student's t fitting. We set $\mu/\vartheta = 10^{-2}$.

since $Z_i = X_i/Y_i = \sqrt{2N_1\varrho\vartheta}\Re\{\tilde{h}_i\}/|\hat{h}_i| \mid |\hat{h}_i| \geq \sqrt{\mu}$. Meanwhile, the characteristic function of T is given by

$$\begin{aligned}\omega_2(t) &\triangleq \mathbb{E}[e^{\imath tT}] = \mathbb{E}\big[e^{\imath t\sum_{i\in\mathcal{K}} Z_i}\big] = \mathbb{E}[e^{\imath tZ}]^K \\ &\overset{(a)}{=} \left[\int_{-\infty}^{\infty} \big(\cos(tz) + \imath\sin(tz)\big) f_Z(z)\mathrm{d}z\right]^K \\ &\overset{(b)}{=} \left[2\int_{0}^{\infty} \cos(tz) f_Z(z)\mathrm{d}z\right]^K, \ \forall t \geq 0, \end{aligned} \tag{27}$$

where (a) comes from exploiting $e^{\imath a} = \cos a + \imath \sin a$ and expressing the expectation in integral form, while (b) is obtained by leveraging the symmetry of $f_Z(s)$ around 0, and properties $\cos(-a) = \cos a$, $\sin(-a) = -\sin a$.

Now, we proceed to match (26) and (27) with the second moment and characteristic function of a scaled Student's t distribution $s\mathcal{T}(\nu)$, which are respectively given by [59]

$$E\big[(s\mathcal{T}(\nu))^2\big] = \frac{s^2\nu}{\nu-2}, \tag{28}$$

$$\mathrm{CF}\big(s\mathcal{T}(\nu)\big) = \frac{K_{\nu/2}(\sqrt{\nu}ts)(\sqrt{\nu}ts)^{\nu/2}}{2^{\nu/2-1}\Gamma(\nu/2)}, \ \forall t \geq 0. \tag{29}$$

Then, the system of equations to solve becomes

$$\left\{\omega_1 = \frac{s^2\nu}{\nu-2}, \qquad \omega_2(t) = \frac{K_{\nu/2}(\sqrt{\nu}ts)(\sqrt{\nu}ts)^{\nu/2}}{2^{\nu/2-1}\Gamma(\nu/2)}\right\}. \tag{30}$$

Through extensive simulation campaigns, we found that the solution of the above system of equations leads to a very accurate fitting when $t = 1$. By setting $t = 1$ and combining the equations in (30), we obtain (22), where ω_2 in (24) matches (27). Then, s is attained from ν by exploiting the first equation in (30). The accuracy of the fitting is illustrated in Fig. 9. □

REFERENCES

[1] N. Mahmood, O. L. A. López, O. Park, I. Moerman, K. Mikhaylov, E. Mercier, A. Munari, F. Clazzer, S. Böcker, and H. Bartz *(Eds.)*, "White paper on critical and massive machine type communication towards 6G," *6G Research Visions*, no. 11, 2020, http://jultika.oulu.fi/files/isbn9789526226781.pdf.

[2] J. Choi, J. Ding, N.-P. Le, and Z. Ding, "Grant-free random access in machine-type communication: Approaches and challenges," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 151–158, 2022.

[3] N. H. Mahmood, O. L. A. López, F. Clazzer, and A. Munari, *Random Access for Cellular Systems*. John Wiley & Sons, Ltd, 2020, pp. 1–25. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119471509.w5GRef030

[4] E. De Carvalho, E. Bjornson, J. H. Sorensen, P. Popovski, and E. G. Larsson, "Random access protocols for massive MIMO," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 216–222, 2017.

[5] O. L. A. López, H. Alves, P. H. J. Nardelli, and M. Latva-aho, "Aggregation and resource scheduling in machine-type communication networks: A stochastic geometry approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 7, pp. 4750–4765, 2018.

[6] O. L. A. Lopez, N. H. Mahmood, H. Alves, C. M. Lima, and M. Latva-aho, "Ultra-low latency, low energy, and massiveness in the 6G era via efficient CSIT-limited scheme," *IEEE Communications Magazine*, vol. 58, no. 11, pp. 56–61, 2020.

[7] O. L. A. López, H. Alves, R. D. Souza, S. Montejo-Sánchez, E. M. G. Fernández, and M. Latva-Aho, "Massive wireless energy transfer: Enabling sustainable IoT toward 6G era," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8816–8835, 2021.

[8] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 2, pp. 264–273, 2013.

[9] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing–the large-scale array regime," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6441–6463, 2013.

[10] L. Ribeiro, M. Leinonen, H. Djelouat, and M. Juntti, "Channel charting for pilot reuse in mMTC with spatially correlated MIMO channels," in *IEEE Globecom Workshops*, 2020, pp. 1–6.

[11] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim, "Compressed sensing for wireless communications: Useful tips and tricks," *IEEE Communications Surveys and Tutorials*, vol. 19, no. 3, pp. 1527–1550, 2017.

[12] O. López, N. Mahmood, M. Shehab, H. Alves, O. Rosabal, L. Marata, and M. Latva-aho, "Statistical tools and methodologies for URLLC–a tutorial," *arXiv preprint arXiv:2212.03292*, 2022.

[13] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 454–465, 2011.

[14] B. Knoop, F. Monsees, C. Bockelmann, D. Wuebben, S. Paul, and A. Dekorsy, "Sparsity-aware successive interference cancellation with practical constraints," in *17th International ITG Workshop on Smart Antennas*, 2013, pp. 1–8.

[15] J. Ahn, B. Shim, and K. B. Lee, "Sparsity-aware ordered successive interference cancellation for massive machine-type communications," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 134–137, 2018.

[16] R. B. Di Renna and R. C. d. Lamare, "Activity-aware multiple feedback SIC for massive machine-type communications," in *12th International ITG Conference on Systems, Communications and Coding*, 2019, pp. 1–6.

[17] H. Djelouat, M. Leinonen, L. Ribeiro, and M. Juntti, "Joint user identification and channel estimation via exploiting spatial channel covariance in mMTC," *IEEE Wireless Communications Letters*, vol. 10, no. 4, pp. 887–891, 2021.

[18] H. Djelouat, M. Leinonen, and M. Juntti, "Spatial correlation aware compressed sensing for user activity detection and channel estimation in massive MTC," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6402–6416, 2022.

[19] ——, "Joint estimation of clustered user activity and correlated channels with unknown covariance in mMTC," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] P. Gao, Z. Liu, P. Xiao, C. H. Foh, and J. Zhang, "Low-complexity block coordinate descend based multiuser detection for uplink grant-free NOMA," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9532–9543, 2022.

[21] H. F. Schepker and A. Dekorsy, "Sparse multi-user detection for CDMA transmission using greedy algorithms," in *8th International Symposium on Wireless Communication Systems*, 2011, pp. 291–295.

[22] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Efficient detectors for joint compressed sensing detection and channel decoding," *IEEE Transactions on Communications*, vol. 63, no. 6, pp. 2249–2260, 2015.

[23] W. Xiong, J. Cao, and S. Li, "Sparse signal recovery with unknown signal sparsity," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–8, 2014.

[24] N. Y. Yu, "A fast and noise-robust algorithm for joint sparse recovery through information transfer," *IEEE Access*, vol. 7, pp. 37 735–37 748, 2019.

[25] K. Lee and N. Y. Yu, "Exploiting prior information for greedy compressed sensing based detection in machine-type communications," *Digital Signal Processing*, vol. 107, p. 102862, 2020.

[26] H. Xiao, W. Chen, J. Fang, B. Ai, and I. J. Wassell, "A grant-free method for massive machine-type communication with backward activity level estimation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6665–6680, 2020.

[27] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," in *42nd Asilomar Conference on Signals, Systems and Computers*, 2008, pp. 581–587.

[28] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1890–1904, 2018.

[29] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6164–6175, 2018.

[30] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Transactions on Signal Processing*, vol. 68, pp. 764–779, 2020.

[31] L. Marata, O. L. A. López, E. N. Tominaga, and H. Alves, "Joint channel estimation and device activity detection in heterogeneous networks," in *29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 836–840.

[32] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2320–2323, 2016.

[33] R. B. Di Renna and R. C. de Lamare, "Joint channel estimation, activity detection and data decoding based on dynamic message-scheduling strategies for mMTC," *IEEE Transactions on Communications*, vol. 70, no. 4, pp. 2464–2479, 2022.

[34] Y. Wang, Y. Wang, T. Wang, and J. Cheng, "Multi-service oriented joint channel estimation and multi-user detection scheme for grant-free massive MTC networks," *IEEE Transactions on Communications*, pp. 1–1, 2023.

[35] M. Ke, Z. Gao, M. Zhou, D. Zheng, D. W. K. Ng, and H. V. Poor, "Next-generation URLLC with massive devices: A unified semi-blind detection framework for sourced and unsourced random access," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 7, pp. 2223–2244, 2023.

[36] N. Ye, J. An, and J. Yu, "Deep-learning-enhanced NOMA transceiver design for massive MTC: Challenges, state of the art, and future directions," *IEEE Wireless Communications*, vol. 28, no. 4, pp. 66–73, 2021.

[37] Y. Bai, B. Ai, and W. Chen, "Deep learning based fast multiuser detection for massive machine-type communication," in *IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–5.

[38] Y. Cui, S. Li, and W. Zhang, "Jointly sparse signal recovery and support recovery via deep learning with applications in MIMO-based grant-free random access," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 788–803, 2021.

[39] W. Chen, B. Zhang, S. Jin, B. Ai, and Z. Zhong, "Solving sparse linear inverse problems in communication systems: A deep learning approach with adaptive depth," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 4–17, 2021.

[40] H. Yu, Z. Fei, Z. Zheng, N. Ye, and Z. Han, "Deep learning-based user activity detection and channel estimation in grant-free NOMA," *IEEE Transactions on Wireless Communications*, vol. 22, no. 4, pp. 2202–2214, 2023.

[41] X. Zhang, F. Labeau, L. Hao, and J. Liu, "Joint active user detection and channel estimation via Bayesian learning approaches in MTC communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 6, pp. 6222–6226, 2021.

[42] X. Zhang, P. Fan, J. Liu, and L. Hao, "Bayesian learning-based multiuser detection for grant-free NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 8, pp. 6317–6328, 2022.

[43] L. Marata, O. Luis Alcaraz López, H. Djelouat, M. Leinonen, H. Alves, and M. Juntti, "Joint coherent and non-coherent detection and decoding techniques for heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1730–1744, 2023.

[44] L. Marata, O. L. A. López, A. Hauptmann, H. Djelouat, and H. Alves, "Joint activity detection and channel estimation for clustered massive machine type communications," *arXiv preprint arXiv:2305.02935*, 2023.

[45] L. Zhao, P. R. Krishnaiah, and Z. Bai, "On detection of the number of signals in presence of white noise," *Journal of multivariate analysis*, vol. 20, no. 1, pp. 1–25, 1986.

[46] L.-C. Zhao, P. Krishnaiah, and Z.-D. Bai, "Remarks on certain criteria for detection of number of signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 2, pp. 129–132, 1987.

[47] W. Chen, K. Wong, and J. Reilly, "Detection of the number of signals: a predicted eigen-threshold approach," *IEEE Transactions on Signal Processing*, vol. 39, no. 5, pp. 1088–1098, 1991.

[48] S. Kritchman and B. Nadler, "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3930–3941, 2009.

[49] M. Chiani and M. Z. Win, "Estimating the number of signals observed by multiple sensors," in *2nd International Workshop on Cognitive Information Processing*, 2010, pp. 156–161.

[50] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory, Volume i*. Prentice-Hall, Inc., 1993.

[51] ——, *Fundamentals of statistical signal processing. detection theory, Volume ii*. Printice Hall PTR, 1998.

[52] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.

[53] X. Lin, "An overview of 5G advanced evolution in 3GPP Release 18," *IEEE Communications Standards Magazine*, vol. 6, no. 3, pp. 77–83, 2022.

[54] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

[55] G. Forchini, "The distribution of the sum of a normal and a t random variable with arbitrary degrees of freedom," *Metron*, vol. 66, no. 2, p. 205, 2008.

[56] R. Marini, K. Mikhaylov, G. Pasolini, and C. Buratti, "Low-power wide-area networks: Comparison of LoRaWAN and NB-IoT performance," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21 051–21 063, 2022.

[57] H. Holma, A. Toskala, and T. Nakamura, *5G technology: 3GPP new radio*. John Wiley & Sons, 2020.

[58] A. Mahmood, M. I. Ashraf, M. Gidlund, and J. Torsner, "Over-the-air time synchronization for URLLC: Requirements, challenges and possible enablers," in *15th International Symposium on Wireless Communication Systems (ISWCS)*, 2018, pp. 1–6.

[59] O. L. Alcaraz López, E. M. Garcia Fernández, and M. Latva-aho, "Fitting the distribution of linear combinations of $t-$variables with more than 2 degrees of freedom," *Journal of Probability and Statistics*, vol. 2023, 2023.