

# On the Impact of Channel Estimation on the Design and Analysis of IRSA based Systems

Chirag Ramesh Srivatsa, *Graduate Student Member, IEEE*,  and Chandra R. Murthy, *Senior Member, IEEE* 

**Abstract**—Irregular repetition slotted aloha (IRSA) is a distributed grant-free random access protocol where users transmit multiple replicas of their packets to a base station (BS). The BS recovers the packets using successive interference cancellation. In this paper, we first derive channel estimates for IRSA, exploiting the sparsity structure of IRSA transmissions, when non-orthogonal pilots are employed across users to facilitate channel estimation at the BS. Allowing for the use of non-orthogonal pilots is important, as the length of orthogonal pilots scales linearly with the total number of devices, leading to prohibitive overhead as the number of devices increases. Next, we present a novel analysis of the throughput of IRSA under practical channel estimation errors, with the use of multiple antennas at the BS. Finally, we theoretically characterize the asymptotic throughput performance of IRSA using a density evolution based analysis. Simulation results underline the importance of accounting for channel estimation errors in analyzing IRSA, which can even lead to 70% loss in performance in severely interference-limited regimes. We also provide novel insights on the effect of parameters such as pilot length, SNR, number of antennas at the BS, etc, on the system throughput.

**Index Terms**—Irregular repetition slotted aloha, pilot contamination, density evolution, channel estimation

## I. INTRODUCTION

Massive machine-type communications (mMTC) is an evolving 5G use-case, expected to serve around  $10^6$  devices per square kilometer [2]. The users in mMTC applications are sporadically active and transmit short packets to a central base station (BS) [3]. Grant-free random access (GFRA) protocols are appropriate in mMTC applications since they incur a low control and signaling overhead [4], [5]. Typically, in these protocols, users transmit packets (consisting of a header containing pilot symbols followed by the data payload) by randomly accessing resource blocks (RBs).<sup>1</sup> Since the length of orthogonal pilots scales linearly with the number of users, the overhead of assigning orthogonal pilots becomes prohibitively expensive. Thus, pilot contamination is inevitable due to the use of non-orthogonal pilots, and has to be accounted for while analyzing the performance of GFRA protocols for mMTC.

Chirag Ramesh Srivatsa is with the Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bangalore, India. Chandra R. Murthy is with the Dept. of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India (e-mail: {chiragramesh, cmurthy}@iisc.ac.in).

This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors. The material includes the proofs of the theorems and lemmas presented in this work. This material is 202133 bytes in size.

This work has been presented in part in [1].

This work was financially supported by research grants from the Qualcomm Innovation Fellowship and the Young Faculty Research Fellowship awarded by the Govt. of India.

<sup>1</sup>We refer to the time-frequency resource as resource blocks (RBs) in this work since each RB can accommodate a whole data packet.

One popular GFRA protocol is irregular repetition slotted aloha (IRSA) [6], [7], which is the focus of this paper. Users in IRSA transmit replicas of their packets on a randomly selected subset of the available RBs. The indices of the RBs in which they transmit make up the access pattern matrix (APM). Existing works in IRSA assume availability of perfect channel state information (CSI) at the BS, which is difficult to achieve, especially when non-orthogonal pilots are employed. Channel estimation errors and pilot contamination due to non-orthogonal pilots can erase much of the gains promised by IRSA protocols. Thus, one of the main goals of this paper is to understand the impact of estimated CSI on the performance of IRSA when non-orthogonal pilots are used.

### A. The IRSA protocol

The decoding in IRSA is an iterative process involving successive interference cancellation (SIC) [8], where the users are decoded via a combination of inter-RB and intra-RB SIC [9]. *Inter-RB SIC* refers to the removal of packet replicas from a different RB than the one the packet was decoded in, while *intra-RB SIC* refers to the removal of a packet from the same RB in which the packet was decoded, in order to facilitate decoding additional packets that may have been transmitted in that RB. The conventional version of IRSA used only inter-RB SIC to decode users and assumed a collision model, wherein only singleton RBs can be decoded [6]. Here, a singleton RB refers to an RB where a single user's packet is received without collision. Since no packets can be decoded in RBs where collisions occur, the maximum possible throughput is one packet per RB, the same as the throughput with perfectly coordinated multiple access. This maximum can be achieved asymptotically as the number of users and RBs go to infinity, when the soliton distribution is used to generate the repetition factors of the users [10].

When the BS is equipped with multiple antennas, it can potentially decode multiple packets in a single RB, i.e., if the signal to interference plus noise ratios (SINRs) of the packets are sufficiently high. Thus, using an *SINR threshold* model has also been considered for IRSA, where users can be decoded if and only if their SINR is higher than a predetermined threshold [11]. After decoding users with sufficiently high SINRs, with a combination of intra-RB and inter-RB SIC, the packet replicas of the decoded users can be removed from all the RBs in which they have transmitted packets. Then, all the RBs can be revisited to see if further users can be decoded from the residual signal. This procedure is continued iteratively until no further packets can be decoded. This yields a higher throughput compared to the collision model, and can potentially achieve a throughput greater than one packet per RB. Thus, a second goal of this paper is to characterize the

performance of IRSAs under estimated CSI as a function of system parameters such as the number of antennas at the BS, the pilot length, the SINR threshold, etc.

### B. Related Works

The throughput of the IRSAs family of multiple access protocols is analyzed using the density evolution (DE) approach, wherein two probability densities are obtained as functions of each other [6]. This iterative recipe provides the asymptotic performance of the system. The asymptotic throughput has been obtained for IRSAs via DE for the MAC [10], accounting for path loss [9], for the scalar Rayleigh fading channel [11], with multiuser detectors [12], for the polarized MIMO channel in satellite networks [13], and other enhanced variants of IRSAs [14], [15]. We have proposed an algorithm to detect the subset of active users in IRSAs [16], wherein we also study the effect of imperfect SIC on IRSAs. In contrast, this work focuses on the effect of channel estimation errors on the performance of IRSAs. A closely related protocol is pattern division multiple access (PDMA) [17], where users transmit their packets across a subset of RBs governed by a binary APM. A difference with PDMA is that the APM is designed in a centralized manner to maximize the so-called *constellation-constrained capacity*, which is not scalable to a massive number of users in mMTC. Thus, a theoretical analysis of the throughput of the IRSAs protocol under pilot contamination, accounting for the effect of channel estimation errors, path loss, fading, and multiple antennas at the BS, is not yet available in the literature, to the best of our knowledge.

### C. Contributions

Our main contributions in this paper are as follows:

- 1) We derive channel estimates for IRSAs under three schemes: the first one exploits the sparsity in the APM to estimate the channels of the users, and the other two assume knowledge of the APM and output minimum mean square error (MMSE) estimates. (See Theorem 1 in Sec. III.)
- 2) We present a novel analysis of the SINR in IRSAs accounting for channel estimation errors, where estimates are acquired via non-orthogonal pilots under the three estimation schemes. (See Theorem 2 in Sec. IV.)
- 3) We theoretically analyze the throughput of IRSAs via DE, when users perform path loss inversion based power control. The analysis reveals the asymptotic performance of the protocols as the number of users and RBs get large. (See Theorem 3 in Sec. V-D and also Sec. V-C.)

Through extensive simulations, we show that channel estimation errors lead to a significant loss of throughput compared to the ideal scenario with perfect CSI at the BS, even resulting in up to 70% loss in severely interference-limited regimes. In particular, in mMTC applications, since it is not possible to assign orthogonal pilots to all users, the resulting pilot contamination can significantly degrade the SINR, leading to poor performance. On the positive side, this loss in performance can be recuperated by optimizing

system parameters such as pilot length, number of antennas, frame length, signal to noise ratio, and SINR threshold. In particular, we show that the pilot length required to obtain near-optimal performance is orders of magnitude lower than the pilot length needed to assign orthogonal pilots to all users. For example, a pilot length of  $\tau = 12$  is sufficient to obtain optimal performance with  $M = 150$  users, whereas the use of orthogonal pilot sequences requires  $\tau = 150$  pilot symbols. (See Fig. 2). This is possible because only a small fraction of users transmit in a given RB in IRSAs; exploiting this sparsity in user access allows one to obtain accurate channel estimates even when the pilots are non-orthogonal. (See Algorithm 1.)

Our analysis also reveals an inflection load, beyond which the system becomes interference-limited, resulting in a dramatic reduction of the throughput. The asymptotic throughput obtained via DE serves as an upper bound for the achievable throughput, and facilitates numerical optimization of the throughput with respect to the system parameters.

*Notation:* The symbols  $a$ ,  $\mathbf{a}$ ,  $\mathbf{A}$ ,  $[\mathbf{A}]_{i,:}$ ,  $[\mathbf{A}]_{:,j}$ ,  $\mathbf{0}_N$ ,  $\mathbf{1}_N$ , and  $\mathbf{I}_N$  denote a scalar, a vector, a matrix, the  $i$ th row of  $\mathbf{A}$ , the  $j$ th column of  $\mathbf{A}$ , all-zero vector of length  $N$ , all ones vector of length  $N$ , and an identity matrix of size  $N \times N$ , respectively.  $[\mathbf{a}]_S$  and  $[\mathbf{A}]_{:,S}$  denote the elements of  $\mathbf{a}$  and the columns of  $\mathbf{A}$  indexed by the set  $S$  respectively.  $\text{diag}(\mathbf{a})$  is a diagonal matrix with diagonal entries given by  $\mathbf{a}$ . The set of real and complex matrices of size  $N \times M$  are denoted as  $\mathbb{R}^{N \times M}$  and  $\mathbb{C}^{N \times M}$ .  $\mathcal{N}(\mathbf{a}, \mathbf{A})$  and  $\mathcal{CN}(\mathbf{a}, \mathbf{A})$  denote the real and complex Gaussian distribution, respectively, with mean  $\mathbf{a}$  and covariance  $\mathbf{A}$ .  $[N]$  denotes the set  $\{1, 2, \dots, N\}$ .  $|\cdot|$ ,  $\|\cdot\|$ ,  $[\cdot]^T$ ,  $[\cdot]^*$ ,  $[\cdot]^H$ ,  $\mathbb{E}[\cdot]$ , and  $\mathbb{E}_{\mathbf{a}}[\cdot]$  denote the magnitude (or cardinality of a set),  $\ell_2$  norm, transpose, conjugate, hermitian, expectation, and the expectation conditioned on  $\mathbf{a}$ , respectively. The superscript  $p$  is used as a descriptive superscript in association with a symbol that is related to the *pilots*. All the other superscripts (or subscripts) that have not been defined as above are indices. A non-exhaustive list of symbols used in this work is presented in Table I.

## II. SYSTEM MODEL

An IRSAs system is considered with  $M$  single-antenna users communicating with a central BS equipped with  $N$  antennas. The users are assumed to be arbitrarily located within a cell, with the BS located at the cell center. The fading is modeled as block-fading, quasi-static and Rayleigh distributed. The time-frequency resource is divided into RBs, and  $T$  RBs together constitute a *frame*. The RBs can be slots, subcarriers or both. In each frame, the users contend for the channel by randomly selecting a subset of RBs, and they transmit replicas of their packets in the selected RBs. Each packet replica comprises of a header containing pilot symbols and payload containing data and error correction symbols.

The access of RBs in a given frame by all the users can be represented by a binary access pattern matrix (APM)  $\mathbf{G} \in \{0, 1\}^{T \times M}$ . The entries of  $\mathbf{G}$  are denoted by  $g_{tm} = [\mathbf{G}]_{tm}$ , and  $g_{tm} = 1$  if the  $m$ th user transmits its packet in the  $t$ th RB, and  $g_{tm} = 0$  otherwise. The  $m$ th user samples their repetition factor  $d_m$  from a preset probability distribution. They then

Table I: Mathematical symbols used in this work.

Symbol	Quantity	Symbol	Quantity	Symbol	Quantity	Symbol	Quantity
$L$	Load	$\gamma_{\text{pr}}$	Threshold used to declare support	$\theta_r$	Success probability	$P$	Data power
$\tau$	Pilot length	$\gamma_{\text{th}}$	Capture threshold	$T$	Number of RBs	$P^{\text{p}}$	Pilot power
$\tau_c$	Packet length	$\mathbf{G}$	Access pattern matrix	$N$	Number of antennas	$N_0$	Noise variance
$\mathcal{T}$	Throughput	$\lambda$	Regularization parameter	$M$	Number of users	$\sigma_{\text{h}}^2$	Channel variance

choose  $d_m$  RBs from a total of  $T$  RBs uniformly at random for transmission. We note that, due to the distributed nature of the protocol, the  $M$  columns of  $\mathbf{G}$  are i.i.d., and  $\mathbf{G}$  is independently generated from one frame to the next.

At the BS, the received signal in the  $t$ th RB is a superposition of the packets transmitted by the users that are scheduled to transmit in the same RB. In the pilot phase, if  $g_{tm} = 1$ , the  $m$ th user transmits a  $\tau$ -length pilot  $\mathbf{p}_m \in \mathbb{C}^\tau$  in the  $t$ th RB, with each pilot symbol transmitted at an average power  $P^{\text{p}}$ , and thus,  $\mathbb{E}[\|\mathbf{p}_m\|^2] = \tau P^{\text{p}}$ . The pilot signal  $\mathbf{Y}_t^{\text{p}} \in \mathbb{C}^{N \times \tau}$  received at the BS using its  $N$  antennas and in the  $t$ th RB is given by

$$\mathbf{Y}_t^{\text{p}} = \sum_{m=1}^M g_{tm} \mathbf{h}_{tm} \mathbf{p}_m^H + \mathbf{N}_t^{\text{p}}, \quad (1)$$

where  $\mathbf{N}_t^{\text{p}} \in \mathbb{C}^{N \times \tau}$  is the complex additive Gaussian noise at the BS with  $[\mathbf{N}_t^{\text{p}}]_{nj} \stackrel{\text{i.i.d.}}{\sim} \mathcal{CN}(0, N_0) \forall n \in [N], j \in [\tau]$  and  $t \in [T]$ , and  $N_0$  is the noise variance. Here  $\mathbf{h}_{tm} = [h_{tm1}, \dots, h_{tmN}]^T$  is the uplink channel vector of the  $m$ th user in the  $t$ th RB, with  $h_{tmn} \stackrel{\text{i.i.d.}}{\sim} \mathcal{CN}(0, \beta_m \sigma_{\text{h}}^2)$ ,  $\forall t \in [T], m \in [M]$  and  $n \in [N]$ , where  $\beta_m$  is the path loss coefficient and  $\sigma_{\text{h}}^2$  is the fading variance.

In the data phase, users transmit their data symbols. Considering one of the data symbols, the  $m$ th user transmits a data symbol  $x_m$  with  $\mathbb{E}[x_m] = 0$  and  $\mathbb{E}[|x_m|^2] = P$ , i.e., with transmit power  $P$ . The corresponding received data signal  $\mathbf{y}_t \in \mathbb{C}^N$  at the BS in the  $t$ th RB is

$$\mathbf{y}_t = \sum_{m=1}^M g_{tm} \mathbf{h}_{tm} x_m + \mathbf{n}_t, \quad (2)$$

where  $\mathbf{n}_t \in \mathbb{C}^N$  is the complex additive white Gaussian noise at the BS with  $[\mathbf{n}_t]_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{CN}(0, N_0), \forall n \in [N]$  and  $t \in [T]$ .

### A. SIC-based Decoding

The received data is processed iteratively at the BS. The BS computes channel estimates for all users in all RBs using the pilot symbols.<sup>2</sup> It uses these channel estimates to combine the received data signal across the BS antennas and attempts to decode the user's data packet, treating interference from other users as noise. If it successfully decodes any user, which can be verified via a cyclic redundancy check, it performs SIC in all RBs in which that user has transmitted, with both inter-RB and intra-RB SIC. The BS proceeds with the next iteration, where the channels are re-estimated for the remaining users, and this decoding process proceeds iteratively.

In this work, we abstract the decoding of a user's packet using an SINR threshold model. That is, if the SINR of a packet in a given RB in any decoding iteration exceeds a

<sup>2</sup>As we will see, when the BS does not know the APM, the BS first detects which users have transmitted in each RB, and computes the channel estimates for the users detected to be active in each of the RBs.

threshold  $\gamma_{\text{th}}$ , then the packet can be decoded correctly [9], [11]. *Packet capture* occurs when a packet can be decoded correctly as per the SINR threshold model, even though it collides with another packet, and is thus considered a good abstraction of the decoding in the physical layer.

We now describe the performance evaluation of IRSa via the SINR threshold model. We first compute channel estimates and SINR achieved by all users in all RBs. If we find a user with  $\text{SINR} \geq \gamma_{\text{th}}$  in some RB, we mark the data packet as having been decoded successfully and remove the contribution of the user's packet from all RBs that contain a replica of that packet. In the next iteration, the channels are re-estimated from the residual pilot symbols after SIC, the SINRs are recomputed in all RBs, and the decoding of users' packets continues. The decoding process proceeds in iterations and stops when no additional users are decoded in two successive iterations. The system throughput  $\mathcal{T}$  is calculated as the number of correctly decoded unique packets divided by the number of RBs.

### B. Overview of the Rest of the Paper

The crucial step in evaluating the performance of the above decoding procedure is the calculation of the SINR at the receiver, which depends on the CSI available at the BS. We first describe the channel estimation process in Sec. III. Then we present the derivation of the SINR in Sec. IV. Finally, we describe the calculation of the asymptotic throughput in Sec. V.

## III. CHANNEL ESTIMATION

In this section, the channel estimates for all users are derived under three schemes. The first scheme, termed the sparsity-based estimation scheme, estimates both the APM and the channels of the users. In contrast with this, the other schemes exploit the knowledge of  $\mathbf{G}$  and output MMSE estimates. This is not a strong assumption and can be made possible by using pseudo-random pattern matrices generated from a seed that is available at the BS and the users.

Channel estimation is performed based on the received pilot signal, which contains the pilots transmitted by all the users who have transmitted in that RB. The estimates are recomputed in every iteration, and hence the signals and channel estimates are indexed by the decoding iteration  $k$ . Let the set of users who have not yet been decoded in the first  $k-1$  iterations be denoted by  $\mathcal{S}_k$ , and for some  $m \in \mathcal{S}_k$ , let  $\mathcal{S}_k^m \triangleq \mathcal{S}_k \setminus \{m\}$ , with  $\mathcal{S}_1 = [M]$ . The received pilot signal at the BS, in the  $t$ th RB, and during the  $k$ th decoding iteration, is given by

$$\mathbf{Y}_t^{pk} = \sum_{i \in \mathcal{S}_k} g_{ti} \mathbf{h}_{ti} \mathbf{p}_i^H + \mathbf{N}_t^{\text{p}}. \quad (3)$$

We now discuss three channel estimation schemes for IRSa.

### A. Sparsity-based APM and Channel Estimation

The first scheme is the sparsity-based estimation scheme in which we estimate the APM and the channels in each decoding iteration. We consider the conjugate transpose of the received pilot signal in the  $t$ th RB from (3) as  $\bar{\mathbf{Y}}_t^{\text{pk}} \triangleq \mathbf{Y}_t^{\text{pk}H}$ , with  $\bar{\mathbf{N}}_t \triangleq \mathbf{N}_t^H$ . Let  $\mathbf{P} \in \mathbb{C}^{\tau \times M}$  contain the known pilots of the  $M$  users as its columns and  $\mathbf{P}^k = [\mathbf{P}]_{:, \mathcal{S}_k}$ . The signal  $\bar{\mathbf{Y}}_t^{\text{pk}}$  can be factorized into the product of two matrices as follows:

$$\bar{\mathbf{Y}}_t^{\text{pk}} = \underbrace{[\mathbf{p}_{i_1}, \dots, \mathbf{p}_{i_{M^k}}]}_{\mathbf{P}^k} \underbrace{\begin{bmatrix} g_{ti_1} \mathbf{h}_{ti_1}^H \\ \vdots \\ g_{ti_{M^k}} \mathbf{h}_{ti_{M^k}}^H \end{bmatrix}}_{\mathbf{Z}_t^k} + \underbrace{\bar{\mathbf{N}}_t}_{\tau \times N}, \quad (4)$$

where  $\mathcal{S}_k = \{i_1, i_2, \dots, i_{M^k}\}$ , with  $M^k = |\mathcal{S}_k|$ . Here,  $\mathbf{Z}_t^k \in \mathbb{C}^{M^k \times N}$  contains the  $t$ th row of the unknown APM  $\mathbf{G}$ , and the unknown channels. The rows of  $\mathbf{Z}_t^k$  are either all-zero or all-nonzero depending on whether the corresponding  $g_{ti} = 0$  or 1. This results in an under-determined system of equations, where the columns of  $\mathbf{Z}_t^k$  share the same support. This structure is called as a multiple measurement vector (MMV) recovery problem in compressed sensing. The estimation of  $\mathbf{Z}_t^k$  from (4) can be performed using well known MMV recovery algorithms from compressed sensing literature to recover  $\{g_{ti}\}$  in the each of the  $T$  RBs.

Multiple sparse Bayesian learning<sup>3</sup> (MSBL) [19] is an empirical Bayesian algorithm that can recover  $\mathbf{Z}_t^k$  from linear under-determined observations  $\bar{\mathbf{Y}}_t^{\text{pk}}$ . In MSBL, a Gaussian prior is imposed on the columns of  $\mathbf{Z}_t^k$  as

$$p(\mathbf{Z}_t^k; \gamma_{kt}) = \prod_{n=1}^N p([\mathbf{Z}_t^k]_{:, n}; \gamma_{kt}) = \prod_{n=1}^N \mathcal{CN}(\mathbf{0}_{M_t}, \Gamma_{kt}), \quad (5)$$

where  $\Gamma_{kt} = \text{diag}(\gamma_{kt})$  and the columns of  $\mathbf{Z}_t^k$  are i.i.d. The elements of  $\gamma_{kt} \in \mathbb{R}_+^{M^k}$  are unknown hyperparameters for the undecoded users. Recovering the hyperparameters would yield  $g_{tm}$  since  $[\gamma_{kt}]_m$  models the variance of the  $m$ th user's channel in the  $t$ th RB. The hyperparameters are estimated by iteratively maximizing the log-likelihood  $\log p(\bar{\mathbf{Y}}_t^{\text{pk}}; \gamma_{kt})$ , with  $p(\bar{\mathbf{Y}}_t^{\text{pk}}; \gamma_{kt}) = \prod_{n=1}^N p([\bar{\mathbf{Y}}_t^{\text{pk}}]_{:, n}; \gamma_{kt})$ .

Let  $j$  denote the iteration index in MSBL. Stated in our notation, the overall estimation procedure is summarized in Algorithm 1. The MSBL algorithm converges to a saddle point or a local maximizer of the overall log-likelihood [19]. Further, the MSBL algorithm has been empirically shown to correctly recover the support of  $\mathbf{Z}_t^k$ , provided  $\tau$  and  $N$  are large enough [19], if the signal to noise ratio is good enough. The algorithm is run for  $j_{\max}$  iterations in each of the  $T$  RBs. As the iterations proceed, the hyperparameters corresponding to users with  $g_{ti} = 0$  converge to zero, resulting in sparse estimates. At the end of the iterations, the estimated coefficient  $\hat{g}_{tm}^k$  for the  $m$ th user in the  $t$ th RB in the  $k$ th decoding iteration is obtained by thresholding  $[\gamma_{kt}^{j_{\max}}]_m$  at a value  $\gamma_{\text{pr}}$ .

<sup>3</sup>Any MMV algorithm can be used to recover joint-sparse columns of  $\mathbf{Z}_t^k$ , but we use MSBL due to its high performance. MSBL also outputs a "plug-in" MMSE channel estimate which can then be used to find a meaningful SINR expression since the estimate is uncorrelated with the estimation error [18].

### Algorithm 1: APM and Channel Estimation in $t$ th RB

---

**Input:**  $\tau, N, N_0, \mathcal{S}_k, \mathbf{P}, \bar{\mathbf{Y}}_t^{\text{pk}}, \gamma_{\text{pr}}, j_{\max}$   
**Compute:**  $M^k = |\mathcal{S}_k|, \mathbf{P}^k = [\mathbf{P}]_{:, \mathcal{S}_k}$   
**Initialize:**  $\gamma_{kt}^0 = \mathbf{1}_{M^k}$   
**for**  $j = 0, 1, 2, \dots, j_{\max}$  **do**  
    Compute  $\Gamma_{kt}^j = \text{diag}(\gamma_{kt}^j)$   
     $\Sigma_{kt}^{j+1} = \Gamma_{kt}^j - \Gamma_{kt}^j \mathbf{P}^{kH} (N_0 \mathbf{I}_\tau + \mathbf{P}^k \Gamma_{kt}^j \mathbf{P}^{kH})^{-1} \mathbf{P}^k \Gamma_{kt}^j$   
     $\mu_{ktN}^{j+1} = N_0^{-1} \Sigma_{kt}^{j+1} \mathbf{P}^{kH} [\bar{\mathbf{Y}}_t^{\text{pk}}]_{:, n}, 1 \leq n \leq N$   
     $[\gamma_{kt}^{j+1}]_i = \frac{1}{N} \sum_{n=1}^N ([\Sigma_{kt}^{j+1}]_{i, n} + |\mu_{ktN}^{j+1}|^2), \forall i \in [M^k]$   
**end**  
**Output:**  $\hat{g}_{tm}^k = \begin{cases} 1, & [\gamma_{kt}^{j_{\max}}]_m \geq \gamma_{\text{pr}}, \forall m \in [M^k], \\ 0, & [\gamma_{kt}^{j_{\max}}]_m < \gamma_{\text{pr}} \end{cases}$ ,  $\hat{\mathbf{Z}}_t^k = [\mu_{kt1}^{j_{\max}} \mu_{kt2}^{j_{\max}} \dots \mu_{ktN}^{j_{\max}}]$

---

This can result in errors in estimating  $g_{ti}$ , and the errors in APM estimation can be described by

$$\mathcal{F}_t^k = \{i \in [M^k] \mid \hat{g}_{ti}^k (1 - g_{ti}) = 1\}, \quad (6a)$$

$$\mathcal{M}_t^k = \{i \in [M^k] \mid (1 - \hat{g}_{ti}^k) g_{ti} = 1\}, \quad (6b)$$

where  $\mathcal{F}_t^k$  is the set of false positive users, and  $\mathcal{M}_t^k$  is the set of false negative users. These errors affect decoding of other users in two ways: both kinds of users contaminate the channel estimates of other users, and users in  $\mathcal{M}_t^k$  interfere with the data decoding of other users as well. The effect of errors in detection of users is described in detail in [16].

The algorithm also outputs the maximum a posteriori probability estimates of the channels  $\hat{\mathbf{Z}}_t^k$  in each of the  $T$  RBs. The estimate  $\hat{\mathbf{H}}_t^k = \hat{\mathbf{Z}}_t^{kH} \in \mathbb{C}^{N \times M_t^k}$  of the channels of the  $M_t^k$  users is described in Theorem 1 and can be calculated as  $\hat{\mathbf{H}}_t^k = \mathbf{Y}_t^{\text{pk}} \mathbf{P}^k \hat{\Gamma}_{kt} (\mathbf{P}^{kH} \mathbf{P}^k \hat{\Gamma}_{kt} + N_0 \mathbf{I}_{M_t^k})^{-1}$ , where  $\hat{\Gamma}_{kt} = \text{diag}(\gamma_{kt}^{j_{\max}})$ . This estimate is a "plug-in" MMSE estimate and it contains estimates for erroneously detected users as well. An added advantage of MSBL is that the path loss coefficient can be calculated by averaging the estimated hyperparameters across RBs as  $\hat{\beta}_i^k = (\sum_{t=1}^T \hat{g}_{ti}^k [\gamma_{kt}^{j_{\max}}]_i) / (\sigma_h^2 \sum_{t=1}^T \hat{g}_{ti}^k)$ . Thus, Algorithm 1 does not require any prior information about the APM or  $\{\beta_i\}_{i=1}^M$  to estimate the channels.

### B. MMSE Channel Estimation with Known APM

We now derive the MMSE channel estimates for all users in each RB, exploiting the knowledge of the APM  $\mathbf{G}$  and  $\{\beta_i\}_{i=1}^M$ . By using a common seed at the BS and the users, the APM can be generated at the BS and thus, we can assume that the BS has knowledge of  $\mathbf{G}$ . Let  $\mathcal{G}_t = \{i \in [M] \mid g_{ti} = 1\}$  be the set of users who have transmitted in the  $t$ th RB. Let  $M_t^k = |\mathcal{G}_t \cap \mathcal{S}_k|$  be the number of users who have transmitted in the  $t$ th RB and have not been decoded in the first  $k-1$  iterations,  $\mathbf{H}_t^k \in \mathbb{C}^{N \times M_t^k}$  denote the channel matrix which contains the channels of the  $M_t^k$  users,  $\mathbf{P}_t^k \in \mathbb{C}^{\tau \times M_t^k}$  denote a matrix that contains the pilot sequences of the  $M_t^k$  users and  $\mathbf{B}_t^k \triangleq \sigma_h^2 \text{diag}(\beta_{i_1}, \beta_{i_2}, \dots, \beta_{i_{M_t^k}})$  be a diagonal matrix containing the path loss coefficients of the  $M_t^k$  users, with  $\mathcal{G}_t \cap \mathcal{S}_k = \{i_1, i_2, \dots, i_{M_t^k}\}$ . Thus, the received signal from

Table II: Channel estimates and error variances under three estimation schemes.

	Sparsity-based estimation with MSBL	MMSE	LCMMSE
$\hat{\mathbf{H}}_t^k$	$\mathbf{Y}_t^{pk} \mathbf{P}_t^k \hat{\mathbf{\Gamma}}_{kt} (\mathbf{P}_t^{kH} \mathbf{P}_t^k \hat{\mathbf{\Gamma}}_{kt} + N_0 \mathbf{I}_{M_t^k})^{-1}$	$\mathbf{Y}_t^{pk} \mathbf{P}_t^k \mathbf{B}_t^k (\mathbf{P}_t^{kH} \mathbf{P}_t^k \mathbf{B}_t^k + N_0 \mathbf{I}_{M_t^k})^{-1}$	$\mathbf{Y}_t^{pk} \mathbf{P}_t^k \text{diag}(\eta_{ti_1}^k, \dots, \eta_{ti_{M_t^k}}^k)$
$\delta_{ti}^k$	$\beta_i \sigma_h^2 \left( \frac{N_0 \ \mathbf{c}_{ti}^k\ ^2 + \sum_{j \in \mathcal{S}_k^i}  r_{jti}^k ^2 \hat{g}_{tj}^k g_{tj} \beta_j \sigma_h^2}{N_0 \ \mathbf{c}_{ti}^k\ ^2 + \sum_{j \in \mathcal{S}_k}  r_{jti}^k ^2 \hat{g}_{tj}^k g_{tj} \beta_j \sigma_h^2} \right)$	$\beta_i \sigma_h^2 \left( \frac{N_0 \ \mathbf{c}_{ti}^k\ ^2 + \sum_{j \in \mathcal{S}_k^i}  r_{jti}^k ^2 g_{tj} \beta_j \sigma_h^2}{N_0 \ \mathbf{c}_{ti}^k\ ^2 + \sum_{j \in \mathcal{S}_k}  r_{jti}^k ^2 g_{tj} \beta_j \sigma_h^2} \right)$	$\beta_i \sigma_h^2 \left( \frac{N_0 \ \mathbf{p}_i\ ^2 + \sum_{j \in \mathcal{S}_k^i}  \mathbf{p}_j^H \mathbf{p}_i ^2 g_{tj} \beta_j \sigma_h^2}{N_0 \ \mathbf{p}_i\ ^2 + \sum_{j \in \mathcal{S}_k}  \mathbf{p}_j^H \mathbf{p}_i ^2 g_{tj} \beta_j \sigma_h^2} \right)$

(3) can be written as  $\mathbf{Y}_t^{pk} = \mathbf{H}_t^k \mathbf{P}_t^{kH} + \mathbf{N}_t^p$ , where  $\mathbf{P}_t^k = [\mathbf{P}]_{:, \mathcal{G}_t \cap \mathcal{S}_k}$ . The MMSE estimate  $\hat{\mathbf{H}}_t^k$  of  $\mathbf{H}_t^k$  is presented in Theorem 1, and can be written as

$$\hat{\mathbf{H}}_t^k = \mathbf{Y}_t^{pk} (\mathbf{P}_t^k \mathbf{B}_t^k \mathbf{P}_t^{kH} + N_0 \mathbf{I}_\tau)^{-1} \mathbf{P}_t^k \mathbf{B}_t^k, \quad (7a)$$

$$\stackrel{(a)}{=} \mathbf{Y}_t^{pk} \mathbf{P}_t^k \mathbf{B}_t^k (\mathbf{P}_t^{kH} \mathbf{P}_t^k \mathbf{B}_t^k + N_0 \mathbf{I}_{M_t^k})^{-1}, \quad (7b)$$

where (a) follows from  $(\mathbf{AB} + \mathbf{I})^{-1} \mathbf{A} = \mathbf{A} (\mathbf{BA} + \mathbf{I})^{-1}$ . Here, the estimate can be calculated via an inverse of either a  $\tau \times \tau$  matrix or an  $M_t^k \times M_t^k$  matrix as required. The MSBL estimate converges to the MMSE estimate when the hyperparameters are estimated well enough, as will be seen in Sec. VI.

### C. Low Complexity MMSE with Known APM

We now describe a low complexity MMSE (LCMMSE) estimate that does not require a matrix inversion computation. For this purpose, the received signal in (3) is right-multiplied by the pilot  $\mathbf{p}_m$  to obtain

$$\mathbf{y}_{tm}^{pk} = \mathbf{Y}_t^{pk} \mathbf{p}_m = \sum_{i \in \mathcal{S}_k} g_{ti} \mathbf{h}_{ti} (\mathbf{p}_i^H \mathbf{p}_m) + \mathbf{N}_t^p \mathbf{p}_m, \quad (8)$$

which is used to find an MMSE estimate of the channel  $\mathbf{h}_{tm}$  of the  $m$ th user in the  $t$ th RB. The LCMMSE channel estimate  $\hat{\mathbf{h}}_{tm}^k$  is described in Theorem 1 and is calculated as

$$\hat{\mathbf{h}}_{tm}^k = \frac{g_{tm} \beta_m \|\mathbf{p}_m\|^2 \sigma_h^2}{N_0 \|\mathbf{p}_m\|^2 + \sum_{i \in \mathcal{S}_k} |\mathbf{p}_i^H \mathbf{p}_m|^2 g_{ti} \beta_i \sigma_h^2} \mathbf{y}_{tm}^{pk} \triangleq \eta_{tm}^k \mathbf{y}_{tm}^{pk}$$

Similar to the MMSE estimate, the LCMMSE estimate uses the knowledge of the APM and  $\{\beta_i\}_{i=1}^M$ . While the MMSE estimator uses the signal  $\mathbf{Y}_t^{pk}$  to compute the estimates, and thus exploits all the information available at the BS, the LCMMSE estimator uses only  $\mathbf{y}_{tm}^{pk}$ , i.e., the projection of  $\mathbf{Y}_t^{pk}$  onto  $\mathbf{p}_m$ , to estimate  $\mathbf{h}_{tm}$ .

The channel estimates under the three schemes and their error variances are given by the following theorem.

**Theorem 1.** *The channel estimate  $\hat{\mathbf{H}}_t^k$  of  $\mathbf{H}_t^k$  in the  $t$ th RB in the  $k$ th decoding iteration, under the three estimation schemes, namely MSBL, MMSE, and LCMMSE, is given in Table II. Specifically, the estimate of the channel  $\mathbf{h}_{ti}$  of the  $i$ th user is calculated as  $\hat{\mathbf{h}}_{ti}^k = [\hat{\mathbf{H}}_t^k]_{:,i}$ . Further, the covariance of the estimation error  $\hat{\mathbf{h}}_{ti}^k \triangleq \hat{\mathbf{h}}_{ti}^k - \mathbf{h}_{ti}$  is  $\delta_{ti}^k \mathbf{I}_N$ , where  $\delta_{ti}^k$  is listed in Table II, with  $\mathbf{c}_t^k = [\mathbf{C}_t^k]_{:,i}$  and  $r_{jti}^k \triangleq \mathbf{p}_j^H \mathbf{c}_t^k$ . For MSBL, we have  $\mathbf{C}_t^k \triangleq \mathbf{P}^k \mathbf{D}_t^k (\mathbf{P}_t^{kH} \mathbf{P}_t^k \mathbf{D}_t^k + N_0 \mathbf{I}_{M_t^k})^{-1}$ , where  $\mathbf{D}_t^k \triangleq \text{diag}(d_{ti_1}^k, d_{ti_2}^k, \dots, d_{ti_{M_t^k}}^k)$  with  $d_{ti}^k = \hat{g}_{ti}^k g_{ti} \beta_i \sigma_h^2$ . For MMSE, we have  $\mathbf{C}_t^k \triangleq \mathbf{P}_t^k \mathbf{B}_t^k (\mathbf{P}_t^{kH} \mathbf{P}_t^k \mathbf{B}_t^k + N_0 \mathbf{I}_{M_t^k})^{-1}$ .*

*Proof.* See supplementary material.  $\blacksquare$

*Remark:* The LCMMSE estimate is composed of two components: a scaling coefficient  $\eta_{tm}^k$  and the post-combined received pilot signal  $\mathbf{y}_{tm}^{pk}$ . From (8), we see that the

received pilot signal  $\mathbf{y}_{tm}^{pk}$  contains pilots of other users, if pilot sequences are not orthogonal. With orthogonal pilots,  $\mathbf{p}_i^H \mathbf{p}_m = 0, \forall i \neq m$ , the LCMMSE estimate is

$$\hat{\mathbf{h}}_{tm}^k = \frac{g_{tm} \beta_m \sigma_h^2}{N_0 + \|\mathbf{p}_m\|^2 g_{tm} \beta_m \sigma_h^2} (g_{tm} \mathbf{h}_{tm} \|\mathbf{p}_m\|^2 + \mathbf{N}_t^p \mathbf{p}_m),$$

and  $\delta_{tm}^k = g_{tm} \beta_m \sigma_h^2 N_0 / (N_0 + g_{tm} \beta_m \sigma_h^2 \|\mathbf{p}_m\|^2)$ , i.e., there is no pilot contamination, and the LCMMSE estimate coincides with the MMSE estimate.

*Complexity:* The MMSE scheme has a complexity of  $\mathcal{O}(\tau^2 M_t^k)$  floating point operations (flops) since it involves inverting a  $\tau \times \tau$  matrix. The MSBL scheme, with  $s$  iterations, has a complexity of  $\mathcal{O}(s\tau^2 M^k)$  flops [19]. The LCMMSE scheme has the lowest complexity of  $\mathcal{O}(M_t^k)$  flops since it does not need any matrix inversion.

## IV. SINR ANALYSIS

In this section, the SINR of each user in all the RBs where it transmits data is derived, accounting for pilot contamination and channel estimation errors. Let  $\rho_{tm}^k$  denote the SINR of the  $m$ th user in the  $t$ th RB in the  $k$ th decoding iteration. Similar to (2), the received data signal in the  $t$ th RB and  $k$ th decoding iteration can be written as

$$\mathbf{y}_t^k = \sum_{i \in \mathcal{S}_k} g_{ti} \mathbf{h}_{ti} x_i + \mathbf{n}_t. \quad (9)$$

A combining vector  $\mathbf{a}_{tm}^k$  is used to decode the  $m$ th user in the  $t$ th RB and  $k$ th decoding iteration, and thus we obtain

$$\begin{aligned} \tilde{\mathbf{y}}_{tm}^k &= \mathbf{a}_{tm}^{kH} \mathbf{y}_t^k = \mathbf{a}_{tm}^{kH} \hat{\mathbf{h}}_{tm}^k g_{tm} x_m - \mathbf{a}_{tm}^{kH} \tilde{\mathbf{h}}_{tm}^k g_{tm} x_m \\ &\quad + \mathbf{a}_{tm}^{kH} \sum_{i \in \mathcal{S}_k^m} g_{ti} \mathbf{h}_{ti} x_i + \mathbf{a}_{tm}^{kH} \mathbf{n}_t, \end{aligned} \quad (10)$$

where  $\tilde{\mathbf{h}}_{tm}^k$  is as defined in Theorem 1. From the above, we see that the signal used to decode the  $m$ th user's data is composed of four terms. The term  $T_1 \triangleq \mathbf{a}_{tm}^{kH} \hat{\mathbf{h}}_{tm}^k g_{tm} x_m$  is the useful signal component of the  $m$ th user; the term  $T_2 \triangleq \mathbf{a}_{tm}^{kH} \tilde{\mathbf{h}}_{tm}^k g_{tm} x_m$  is contributed by the channel estimation error  $\tilde{\mathbf{h}}_{tm}^k$  of the  $m$ th user; the term  $T_3 \triangleq \sum_{i \in \mathcal{S}_k^m} \mathbf{a}_{tm}^{kH} \mathbf{h}_{ti} g_{ti} x_i$  captures the inter-user interference from the users who have also transmitted in the  $t$ th RB and have not yet been decoded upto the  $k$ th decoding iteration; and the last term  $T_4 \triangleq \mathbf{a}_{tm}^{kH} \mathbf{n}_t$  is the additive noise component.

In order to compute the SINR, the power in the received signal is calculated conditioned on the knowledge of the estimates [20]. Since MMSE estimates are employed, all three estimates are uncorrelated with the channel estimation error, and thus  $T_2$  is uncorrelated with  $T_1$ . The additive noise is uncorrelated with the signal, and since the users' data signals are independent,  $T_3$  is uncorrelated with the other terms. Thus, all four components in the received signal are uncorrelated

and the total power is the sum of the powers of the individual components

$$\mathbb{E}_{\mathbf{z}}[|\tilde{y}_{tm}^k|^2] = \sum_{i=1}^4 \mathbb{E}_{\mathbf{z}}[|T_i|^2], \quad (11)$$

where  $\mathbf{z}$  contains the channel estimates of the users. The SINR for all the users is now presented.

**Theorem 2.** *The signal to interference plus noise ratio (SINR) achieved by the  $m$ th user in the  $t$ th RB in the  $k$ th decoding iteration can be written as*

$$\rho_{tm}^k = \frac{\text{Gain}_{tm}^k}{N_0/P + \text{MUI}_{tm}^k + \text{Est}_{tm}^k}, \quad \forall m \in \mathcal{S}_k, \quad (12)$$

where  $\text{Gain}_{tm}^k$  represents the useful signal power of the  $m$ th user,  $\text{MUI}_{tm}^k$  represents the multi-user interference power of other users, and  $\text{Est}_{tm}^k$  represents the interference power caused due to the channel estimation errors. Under MMSE and LCMMSE channel estimation, these can be expressed as

$$\begin{aligned} \text{Gain}_{tm}^k &= g_{tm} \frac{|\mathbf{a}_{tm}^{kH} \hat{\mathbf{h}}_{tm}^k|^2}{\|\mathbf{a}_{tm}^k\|^2}, \quad \text{MUI}_{tm}^k = \sum_{i \in \mathcal{S}_k^m} g_{ti} \frac{|\mathbf{a}_{tm}^{kH} \hat{\mathbf{h}}_{ti}^k|^2}{\|\mathbf{a}_{tm}^k\|^2}, \\ \text{Est}_{tm}^k &= \sum_{i \in \mathcal{S}_k} g_{ti} \delta_{ti}^k. \end{aligned}$$

With the sparsity-based scheme, the SINR denominator contains an additional term,  $\text{FNU}_{tm}^k$ , which represents the interference power caused due to false negative users. The corresponding terms with MSBL can be expressed as

$$\begin{aligned} \text{Gain}_{tm}^k &= \hat{g}_{tm}^k g_{tm} \frac{|\mathbf{a}_{tm}^{kH} \hat{\mathbf{h}}_{tm}^k|^2}{\|\mathbf{a}_{tm}^k\|^2}, \quad \text{MUI}_{tm}^k = \sum_{i \in \mathcal{S}_k^m} \hat{g}_{ti}^k g_{ti} \frac{|\mathbf{a}_{tm}^{kH} \hat{\mathbf{h}}_{ti}^k|^2}{\|\mathbf{a}_{tm}^k\|^2}, \\ \text{Est}_{tm}^k &= \sum_{i \in \mathcal{S}_k} \hat{g}_{ti}^k g_{ti} \delta_{ti}^k, \quad \text{FNU}_{tm}^k = \sum_{i \in \mathcal{S}_k^m} (1 - \hat{g}_{ti}^k) g_{ti} \beta_i \sigma_n^2. \end{aligned}$$

Here, the estimates  $\hat{\mathbf{h}}_{ti}^k = [\hat{\mathbf{H}}_t^k]_{:,i}$  and the error variances  $\delta_{ti}^k$  are obtained from Theorem 1 for all the three schemes.

*Proof.* See supplementary material.  $\blacksquare$

The SINR expression derived in Theorem 2 is applicable to any arbitrary receive combining scheme given by the matrix  $\mathbf{A}_t^k$ , with  $\mathbf{a}_{tm}^k = [\mathbf{A}_t^k]_{:,m}$ . When regularized zero forcing (RZF) combining is used, the combining matrix is

$$\mathbf{A}_t^k = \hat{\mathbf{H}}_t^k (\hat{\mathbf{H}}_t^{kH} \hat{\mathbf{H}}_t^k + \lambda \mathbf{I}_{M_t^k})^{-1}, \quad (13)$$

where  $\lambda$  is the regularization parameter. The SINR with RZF can be computed by substituting the columns of the above matrix into (12). We now describe two popular combining schemes, which are special cases of RZF, in which simpler expressions for the SINR can be computed.<sup>4</sup> The expressions are written for MMSE/LCMMSE, and can be extended to MSBL as detailed in Theorem 2.

1) *Maximal Ratio Combining (MRC):* MRC is obtained from RZF as  $\lambda \rightarrow \infty$  and the combining matrix becomes  $\mathbf{A}_t^k = \hat{\mathbf{H}}_t^k$ . Thus  $\mathbf{a}_{tm}^k = \hat{\mathbf{h}}_{tm}^k$ , and SINR can be computed as

$$\rho_{tm}^k = \frac{P g_{tm} \|\hat{\mathbf{h}}_{tm}^k\|^2}{N_0 + \sum_{i \in \mathcal{S}_k} P g_{ti} \delta_{ti}^k + \sum_{i \in \mathcal{S}_k^m} P g_{ti} \frac{|\hat{\mathbf{h}}_{tm}^{kH} \hat{\mathbf{h}}_{ti}^k|^2}{\|\hat{\mathbf{h}}_{tm}^k\|^2}}. \quad (14)$$

<sup>4</sup>In this paper, we do not consider the MMSE combiner, which is a special case of RZF combining [20].

---

## Algorithm 2: Performance Evaluation of IRSAs

---

```

Input:  $\tau, N, T, M, N_0, \mathbf{G}, \mathbf{P}, \{\mathbf{Y}_t^p\}_{t=1}^T, k_{\max}$ 
Initialize:  $\mathcal{S}_1 = [M], \mathcal{G}_t = \{i \in [M] | g_{ti} = 1\}$ 
for  $k = 1, 2, 3, \dots, k_{\max}$  do
    for  $t = 1, 2, \dots, T$  do
        Find  $M_t^k = |\mathcal{G}_t \cap \mathcal{S}_k|, \mathbf{P}_t^k = [\mathbf{P}]_{:, \mathcal{G}_t \cap \mathcal{S}_k}, \mathbf{Y}_t^{pk}$ 
        Compute  $\hat{\mathbf{h}}_{ti}^k, \forall i \in \mathcal{S}_k$  via Theorem 1
        Evaluate the SINR  $\rho_{ti}^k$  via Theorem 2
        If  $\rho_{ti}^k \geq \gamma_{\text{th}}$ , remove user  $i$  from  $\mathcal{S}_k$  and
            perform IC in all RBs where  $g_{ti} = 1$ 
    end
end
Output:  $\mathcal{T} = (M - |\mathcal{S}_{k_{\max}}|)/T, \text{PLR} = |\mathcal{S}_{k_{\max}}|/M$ 

```

---

2) *Zero Forcing (ZF):* The RZF combiner reduces to the ZF combiner as  $\lambda \rightarrow 0$ . The inverse of the gram-matrix of the channel estimates exists with probability one when  $N \geq M_t^k$  and  $\hat{\mathbf{H}}_t^k$  has full column rank.<sup>5</sup> Hence, we can compute the combining matrix as  $\mathbf{A}_t^k = \hat{\mathbf{H}}_t^k (\hat{\mathbf{H}}_t^{kH} \hat{\mathbf{H}}_t^k)^{-1}$ . Using the above, it is easy to show that the SINR expression simplifies as [20]

$$\rho_{tm}^k = \frac{P g_{tm}}{(N_0 + \sum_{i \in \mathcal{S}_k} P g_{ti} \delta_{ti}^k) [(\hat{\mathbf{H}}_t^{kH} \hat{\mathbf{H}}_t^k)^{-1}]_{mm}}. \quad (15)$$

Note that the third term in the denominator of (14) has been suppressed with ZF combining. However, due to pilot contamination, the term  $[(\hat{\mathbf{H}}_t^{kH} \hat{\mathbf{H}}_t^k)^{-1}]_{mm}$  may contain contributions from the channels of all users. As a consequence, the gram matrix could be ill-conditioned, and the denominator term could be large. Thus, the pilot length, which determines the pilot contamination incurred, is crucial in comparing the performance obtained by the combining schemes. The system throughput can now be calculated from the above SINR expressions via the decoding model described in Sec. II-A, and is described in Algorithm 2 for MMSE/LCMMSE. For MSBL, the initial step in each RB instead consists of finding  $M^k = |\mathcal{S}_k|$ , and  $\mathbf{P}^k = [\mathbf{P}]_{:, \mathcal{S}_k}$ . We also estimate  $\{g_{ti}\}$  and  $\{\mathbf{h}_{ti}\}$  via Algorithm 1 before finding the SINR.

Before proceeding with the analysis of the throughput, we briefly discuss the SINR in the massive MIMO regime, which helps us in interpreting the SINR expressions. We note that the results presented in Sec. VI hold true for any  $N$ . However, when  $N$  is large, a simpler expression for SINR with MRC can be obtained as follows.

**Lemma 1.** *As the number of antennas  $N$  gets large, the SINR with MRC converges almost surely to*

$$\bar{\rho}_{tm}^k = \frac{N \text{Sig}_{tm}^k}{\epsilon_{tm}^k (N_0/P + \text{IntNC}_{tm}^k) + \text{IntC}_{tm}^k}, \quad (16)$$

where  $\text{Sig}_{tm}^k$  is the desired signal,  $\text{IntNC}_{tm}^k$  represents the non-coherent interference, and  $\text{IntC}_{tm}^k$  represents the coherent interference. Each of these can be found in Table III. Here,  $\delta_{tm}^k$

<sup>5</sup>We note that the condition  $N \geq M_t^k$  is not hard to satisfy in IRSAs. For example, with  $L = 2, 3, 4$ , each RB will be occupied by 6, 9, 12 users on an average, respectively, if the average repetition factor is  $\bar{d} = 3$ . Thus any  $N$  greater than, say, 16 would be sufficient to decode the users in most RBs.

Table III: Deterministic equivalent approximation to the SINR.

	Sparsity-based estimation with MSBL	MMSE	LCMMSE
$\epsilon_{tm}^k$	$N_0 \ \mathbf{c}_{tm}^k\ ^2 + \sum_{i \in \mathcal{S}_k} g_{ti} \beta_i \sigma_h^2  \mathbf{c}_{tm}^{kH} \mathbf{p}_i ^2$	$N_0 \ \mathbf{c}_{tm}^k\ ^2 + \sum_{i \in \mathcal{S}_k} g_{ti} \beta_i \sigma_h^2  \mathbf{c}_{tm}^{kH} \mathbf{p}_i ^2$	$N_0 \ \mathbf{p}_m\ ^2 + \sum_{i \in \mathcal{S}_k} g_{ti} \beta_i \sigma_h^2  \mathbf{p}_m^H \mathbf{p}_i ^2$
$\text{Sig}_{tm}^k$	$\hat{g}_{tm}^k g_{tm} (\epsilon_{tm}^k)^2$	$g_{tm} (\epsilon_{tm}^k)^2$	$g_{tm} \beta_m^2 \sigma_h^4 \ \mathbf{p}_m\ ^4$
$\text{IntNC}_{tm}^k$	$\hat{g}_{tm}^k g_{tm} \delta_{tm}^k + \sum_{i \in \mathcal{S}_k^m} g_{ti} \beta_i \sigma_h^2$	$g_{tm} \delta_{tm}^k + \sum_{i \in \mathcal{S}_k^m} g_{ti} \beta_i \sigma_h^2$	$g_{tm} \delta_{tm}^k + \sum_{i \in \mathcal{S}_k^m} g_{ti} \beta_i \sigma_h^2$
$\text{IntC}_{tm}^k$	$N \sum_{i \in \mathcal{S}_k^m} g_{ti} \beta_i^2 \sigma_h^4  \mathbf{c}_{tm}^{kH} \mathbf{p}_i ^2$	$N \sum_{i \in \mathcal{S}_k^m} g_{ti} \beta_i^2 \sigma_h^4  \mathbf{c}_{tm}^{kH} \mathbf{p}_i ^2$	$N \sum_{i \in \mathcal{S}_k^m} g_{ti} \beta_i^2 \sigma_h^4  \mathbf{p}_m^H \mathbf{p}_i ^2$

and  $\mathbf{c}_{tm}^k$  are obtained from Theorems 1 and 2, respectively, for the three estimation schemes.

*Proof.* See supplementary material.  $\blacksquare$

*Remark:*  $\text{IntNC}_{tm}^k$  arises due to channel estimation errors and is independent of  $N$ , while  $\text{IntC}_{tm}^k$  is due to pilot contamination and increases linearly with  $N$ . Further, since  $\bar{\rho}_{tm}^k$  is independent of the fading states of each user, it assures successful recovery of packets with high probability if  $\bar{\rho}_{tm}^k \gg \gamma_{\text{th}}$ . Similarly, the packet will not be decodable with probability close to 1 if  $\bar{\rho}_{tm}^k \ll \gamma_{\text{th}}$ . However, it turns out that in order to characterize the throughput of IRSA, it is necessary to capture the statistics of the SINR when  $\bar{\rho}_{tm}^k \approx \gamma_{\text{th}}$ . The small fluctuations in  $\rho_{tm}^k$  around  $\bar{\rho}_{tm}^k$  due to fading, and the resulting probability of packet decoding error, need to be calculated accurately. Hence, the calculation of the statistics of the SINR using (12) is vital to find the throughput of IRSA. We address this in the next section.

## V. THEORETICAL ANALYSIS OF THROUGHPUT

Density Evolution (DE) analysis has been applied to characterize the asymptotic performance of message passing-based decoding on graphs for low density parity check codes [21] and IRSA [6]. In this section, the representation of IRSA decoding as a bipartite graph is discussed first. Then the graph perspective distributions are defined, the failure probabilities are derived, and finally, the asymptotic throughput of IRSA is characterized. It is assumed that users perform path loss inversion-based power control. We note that a closed form expression for the throughput cannot be derived even for the most basic variant of IRSA due to the underlying graph structure [6]. Hence, we need to resort to DE, which provides an iterative recipe to compute the throughput.

SIC-based decoding can be viewed as message passing on a bipartite graph [6], and thus IRSA, which uses SIC decoding, can be decoded on graphs. A typical IRSA frame can be represented as a bipartite graph, which is made up of  $M$  user nodes (one node for each user),  $T$  RB nodes (one node for each RB), and the edges between them. An edge connects a user node to an RB node if and only if that user has transmitted a packet in that corresponding RB. For example, in Fig. 1, there will be an edge between user node  $u_1$  and RB node  $s_1$  if and only if user  $u_1$  has transmitted a packet replica in RB  $s_1$ . During decoding, edges that connect to users whose SINR is above a threshold are removed from each RB. Each decoding iteration consists of several intra-RB SIC and inter-RB SIC steps. Once an SIC step is performed, the corresponding edge in the bipartite graph is removed. Thus, the edge between user node  $u_1$  and RB node  $s_1$  is removed if the user  $u_1$  is decoded

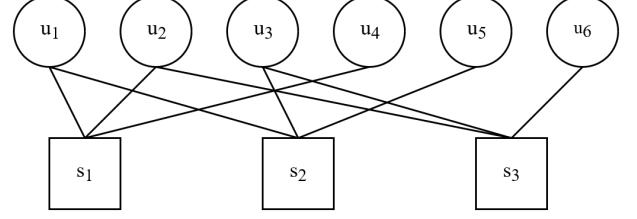


Fig. 1: IRSA represented as a bipartite graph.

in any of the RBs in which the user has transmitted a packet. Decoding is successful if, at the end of the SIC process, all edges in the graph get removed. A decoding failure is declared if not all edges have been removed or no new edge is removed from the graph in two consecutive iterations.

### A. Graph Perspective Degree Distributions

The total number of packets transmitted by a user in a given frame is referred to as the repetition factor of that user. It is equal to the degree of the user node at the start of decoding, and is the same as the number of edges connected to that user node in the bipartite graph representation of SIC decoding. The *node-perspective user degree distribution* is defined as the set of probabilities  $\{\phi_d\}_{d=2}^{d_{\max}}$ , where  $\phi_d$  represents the probability that a user has a repetition factor  $d$  with  $d_{\max}$  being the maximum number of RBs in which any user is allowed to transmit. Here,  $\phi_d$  is nonzero for  $d \geq 2$  since each user transmits at least 2 packets in IRSA.

The total number of packets received in an RB is referred to as the collision factor of that RB. It is equal to the degree of the RB node at the start of decoding, which is the number of edges connected to that RB node. The *node-perspective RB degree distribution* is defined as the set of probabilities  $\{\psi_c\}_{c=0}^M$ , where  $\psi_c$  represents the probability that an RB has a collision factor  $c$ . The polynomial representations of the node-perspective user and RB degree distributions are

$$\phi(x) = \sum_{d=2}^{d_{\max}} \phi_d x^d, \quad \psi(x) = \sum_{c=0}^M \psi_c x^c, \quad (17)$$

respectively. The corresponding *edge-perspective user and RB degree distributions* are defined as  $\lambda(x) = \sum_{d=2}^{d_{\max}} \lambda_d x^{d-1} = \phi'(x)/\phi'(1)$ ;  $\xi(x) = \sum_{c=1}^M \xi_c x^{c-1} = \psi'(x)/\psi'(1)$ , respectively, where  $\lambda_d = d\phi_d/\phi'(1)$  represents the probability that an edge is connected to a user with repetition factor  $d$  and  $\xi_c = c\psi_c/\psi'(1)$  represents the probability that an edge is connected to an RB with collision factor  $c$ .

The *input load*  $L$  of the system is defined as the ratio of the number of users to the number of RBs,  $L \triangleq M/T$ . The average repetition factor is  $\bar{d} = \phi'(1) = \sum_d d\phi_d$  and the average collision factor is  $\bar{c} = \psi'(1) = \sum_c c\psi_c$ , making the

load  $L = M/T = \bar{c}/\bar{d}$ . Since  $\bar{c} = L\bar{d}$ , fixing the load and the node-perspective user degree distribution fixes the other three degree distributions as well. The probability that a generic user, from a total of  $M$  users, transmits within an RB is  $\bar{c}/M$ . Since the users transmit their packets independently of each other,  $\psi_c$  follows a binomial distribution. Thus, the coefficients of the polynomials representing the node and edge-perspective RB degree distributions are respectively given by

$$\psi_c = \binom{M}{c} \left(\frac{\bar{c}}{M}\right)^c \left(1 - \frac{\bar{c}}{M}\right)^{M-c}, \quad (18a)$$

$$\text{and } \xi_c = \binom{M-1}{c-1} \left(\frac{\bar{c}}{M}\right)^{c-1} \left(1 - \frac{\bar{c}}{M}\right)^{M-c}. \quad (18b)$$

For a fixed  $L = M/T$ , as  $M, T \rightarrow \infty$ , the node-perspective and edge-perspective RB degree distributions, which are binomial, become Poisson distributed [22]:

$$\psi_c = \frac{(\bar{c})^c \exp(-\bar{c})}{c!} \text{ and } \xi_c = \frac{(\bar{c})^{c-1} \exp(-\bar{c})}{(c-1)!}. \quad (19)$$

We now use the degree distributions defined above to find the failure probabilities in the next subsection.

### B. Failure Probabilities

In the case of a decoding failure, failure messages are exchanged along the edges between the user and the RB nodes. The probability that an edge carries a failure message from an RB node to a user node in the  $i$ th iteration is denoted by  $p_i$ . The probability that an edge carries a failure message from a user node to an RB node in the  $i$ th iteration is denoted by  $q_i$ .

The failure probability  $q_i$  is calculated using the edge-perspective user degree distribution as

$$q_i = \sum_{d=2}^{d_{\max}} \lambda_d q_i^{(d)} = \sum_{d=2}^{d_{\max}} \lambda_d p_{i-1}^{d-1} = \lambda(p_{i-1}). \quad (20)$$

Here,  $q_i^{(d)}$  is the probability that an edge carries a failure message in the  $i$ th iteration given that it is connected to a user node with repetition factor  $d$ . The edges carry a failure message from a user if and only if all the other  $d-1$  incoming edges to that user carry failure messages in the previous iteration, i.e.,  $q_i^{(d)} = p_{i-1}^{d-1}$ .

The failure probability  $p_i$  is calculated using the edge-perspective RB degree distribution as

$$p_i = \sum_{c=1}^M \xi_c p_i^{(c)} \xrightarrow{M \rightarrow \infty} p_i = \sum_{c=1}^{\infty} \xi_c p_i^{(c)}, \quad (21)$$

where  $p_i^{(c)}$  is the probability that an edge carries a failure message in the  $i$ th iteration given that it is connected to an RB node with collision factor  $c$ . DE is applicable as  $M$  and  $T \rightarrow \infty$  with  $L = M/T$  kept fixed [6]. Hence the above probability is computed as an infinite summation.

In the SINR threshold model, decoding failure happens at an RB node if the SINR of all users who have transmitted in that RB and have not yet been decoded is below the SINR threshold. This constitutes a failure message from the RB node [11]. In order to determine  $p_i^{(c)}$ , any one of the  $c$  packets is considered to be a reference packet, which can get decoded

with a combination of intra-RB and inter-RB SIC. Separating the intra-RB and inter-RB SIC,  $p_i^{(c)}$  can be evaluated as

$$p_i^{(c)} = 1 - \sum_{r=1}^c \theta_r \binom{c-1}{r-1} q_i^{r-1} (1 - q_i)^{c-r}. \quad (22)$$

Here,  $\theta_r$  denotes the probability that the reference packet gets decoded in the current decoding iteration starting from degree  $r$  using only intra-RB SIC, and  $\binom{c-1}{r-1} q_i^{r-1} (1 - q_i)^{c-r}$  denotes the probability that the collision factor of the RB node reduces from  $c$  to  $r$  using only inter-RB SIC [9]. The evaluation of  $\theta_r$  is discussed in Sec. V-D. Substituting for  $p_i^{(c)}$  from (22), we obtain  $p_i$  as a function of  $q_i$ :

$$p_i = 1 - \sum_{c=1}^{\infty} \sum_{r=1}^c \xi_c \theta_r \binom{c-1}{r-1} q_i^{r-1} (1 - q_i)^{c-r}. \quad (23)$$

Thus, we compute the failure probabilities  $p_i$  and  $q_i$  recursively from each other, as observed in (20) and (23).

### C. Evaluation of Throughput

We now describe the evaluation of the throughput. Substituting for  $\xi_c$  from (19), we can simplify (23) to

$$p_i = 1 - e^{-\bar{c}q_i} \sum_{r=1}^{\infty} \theta_r \frac{(\bar{c}q_i)^{r-1}}{(r-1)!} \triangleq f(q_i). \quad (24)$$

Thus,  $q_i = \lambda(p_{i-1})$  and  $p_i = f(q_i)$  are calculated alternately as functions of each other as seen in (20) and (24). The procedure can be initialized with either  $q_0 = 1$  or  $p_0 = f(1)$ .

The failure probability at the end of decoding is  $p_{\infty} = \lim_{i \rightarrow \infty} p_i$  and  $(p_{\infty})^d$  is the probability that a packet transmitted from a user with repetition factor  $d$  does not get decoded at the receiver. Therefore, the asymptotic packet loss rate (PLR), which is the fraction of packets that are not decoded at the BS, is calculated as

$$\text{PLR} = \phi(p_{\infty}) = \sum_{d=2}^{d_{\max}} \phi_d(p_{\infty})^d. \quad (25)$$

The asymptotic throughput of the system can now be obtained from the asymptotic PLR as<sup>6</sup>

$$\mathcal{T} = L(1 - \text{PLR}). \quad (26)$$

The iterations  $p_i = f(\lambda(p_{i-1}))$  converge asymptotically to  $p_{\infty} = 0$  if the system load  $L < L^*$  [6]. Here,  $L^*$  is called the *inflection load* of the system: for any  $L \geq L^*$ , the system becomes interference limited and the PLR does not converge to 0 as  $L$  increases. Thus, for  $L < L^*$ ,  $p_{\infty} = 0$  and therefore the asymptotic PLR = 0, and the throughput equals  $L$ . For  $L \geq L^*$ , the throughput decreases monotonically with  $L$ .

The crucial step in the evaluation of the throughput lies in the computation of  $\theta_r$ , which we now describe.

### D. Characterization of $\theta_r$

We now describe a procedure to evaluate the success probability  $\theta_r$ , which is the probability of decoding the reference packet in an RB with degree  $r$  via intra-RB SIC only. There are  $r$  users whose packets have not yet been decoded in the RB. The reference packet can get decoded in any of the

<sup>6</sup>The DE process yields an iterative recipe to obtain the asymptotic throughput and cannot be used to analytically find a relationship between the system parameters and the throughput.

intra-RB SIC steps. The packets with SINR higher than that of the reference packet get decoded first. Further, the reference packet can only be decoded if decoding has been successful for higher SINR packets, i.e., if they satisfied  $\text{SINR} \geq \gamma_{\text{th}}$  as well. Thus,  $\theta_r$  is the joint probability that the reference packet and the packets with higher SINRs all get decoded.

Clazzer et al. [11] evaluate  $\theta_r$  as the probability " $D(r)$ " under a Rayleigh fading SISO channel setup with a perfect CSI assumption. The same method cannot be applied here, since we consider MIMO Rayleigh fading and account for imperfect CSI due to pilot contamination and channel estimation errors. In particular, in a MIMO setup, it is possible that multiple users' SINR simultaneously exceed the decoding threshold. Further, their work is limited to the case where the decoding threshold  $\gamma_{\text{th}}$  is such that only one user can be decoded in any decoding iteration, while we make no such assumptions.

Since  $\theta_r$  is evaluated based on the SINR of multiple users in a *single RB*, we consider only one RB wherein  $r$  users have transmitted their packets. These users are decoded via only intra-RB iterations since there is only a single RB under consideration. Let the set of users who have not yet been decoded in the first  $k-1$  intra-RB decoding iterations be denoted by  $\mathcal{S}_k$ , and  $\mathcal{S}_k^m \triangleq \mathcal{S}_k \setminus \{m\}$ , with  $\mathcal{S}_1 = [r]$ .<sup>7</sup> In each intra-RB decoding iteration, a single user with the highest SINR is decoded if their SINR  $\geq \gamma_{\text{th}}$ .

The SINR of the  $m$ th user in the  $k$ th intra-RB decoding iteration,  $\rho_m^k$ , is calculated as seen before in Theorem 2. Specifically, when users are only decoded via intra-RB SIC within one RB, we obtain the SINR as

$$\rho_m^k = \frac{|\mathbf{a}_m^{kH} \hat{\mathbf{h}}_m^k|^2}{\|\mathbf{a}_m^k\|^2 (N_0/P + \sum_{i \in \mathcal{S}_k} \delta_i^k) + \sum_{i \in \mathcal{S}_k^m} |\mathbf{a}_m^{kH} \hat{\mathbf{h}}_i^k|^2}. \quad (27)$$

Here,  $\delta_i^k$  is the error variance of the  $i$ th user in the  $k$ th intra-RB decoding iteration,  $\hat{\mathbf{h}}_m^k$  is the channel estimate of the  $m$ th user, both obtained from Theorem 1, and  $\mathbf{a}_m^k$  is the combining vector for the  $m$ th user.<sup>8</sup> Let  $\rho_{\text{max}}^k$  denote the SINR of the user with the highest SINR in the  $k$ th intra-RB decoding iteration, calculated as  $\rho_{\text{max}}^k = \max_{m \in \mathcal{S}_k} \rho_m^k$ . Let  $s$  be the index of the intra-RB decoding iteration in which the reference packet is decoded, with  $1 \leq s \leq r$ . Thus,  $\theta_r$  is calculated as

$$\theta_r = \Pr(\rho_{\text{max}}^1 \geq \gamma_{\text{th}}, \rho_{\text{max}}^2 \geq \gamma_{\text{th}}, \dots, \rho_{\text{max}}^s \geq \gamma_{\text{th}}). \quad (28)$$

Recall that the reference packet is tagged uniformly at random from the users. With path loss inversion based power control, users have identical channel statistics, and thus,  $\theta_r$  is independent of which packet is tagged as the reference packet.

The computation of the success probability  $\theta_r$  is involved because there is no clear relation between the peak SINRs across decoding iterations. Also, the channel estimates of different users are correlated, across both the user index and the decoding iteration index, making it difficult to use order statistics. Further,  $\theta_r$  is dependent on a large number of

<sup>7</sup>The set  $\mathcal{S}_k$  as defined here is a slight abuse of notation. In Sec. III, the set  $\mathcal{S}_k$  consisted of users being decoded via both intra-RB and inter-RB iterations, whereas here,  $\mathcal{S}_k$  consists of users being decoded via only intra-RB iterations.

<sup>8</sup>Since the decoding process with intra-RB SIC involves only the RB in consideration, the RB index and the APM are dropped in this section.

random channel vectors, the order statistics of the peak SINRs, and the pilot sequences of all the users. As a consequence,  $\theta_r$  cannot be found in closed form, and needs to be empirically evaluated. However, we present three approximations to  $\theta_r$ , which are valid when perfect CSI is available at the BS, i.e., there is no pilot contamination or estimation errors. The assumptions are made for analytical tractability. These lead to interpretable expressions for the SINR and  $\theta_r$ , and provide upper bounds on the throughput with estimated CSI.

**Theorem 3.** *When perfect CSI is available at the BS, and MRC is used for decoding,  $\theta_1$  is given by*

$$\theta_1 = \Gamma_{\text{inc}}(N, \rho_0^{-1} \gamma_{\text{th}}) / \Gamma(N), \quad (29)$$

where  $\rho_0 \triangleq P\sigma_h^2/N_0$ ,  $\Gamma_{\text{inc}}(s, x) = \int_x^\infty t^{s-1} e^{-t} dt$  is the upper incomplete gamma function, and  $\Gamma(s)$  is the ordinary gamma function. For  $r \geq 2$ , the SINR with MRC and large  $N$  can be computed as  $\rho_m^k = N(\rho_0^{-1} + N \sum_{i \in \mathcal{S}_k^m} t_{mi})^{-1}$ , where  $t_{mi} \triangleq |\mathbf{h}_m^H \mathbf{h}_i|^2 / (\|\mathbf{h}_m\|^2 \|\mathbf{h}_i\|^2)$ . With  $t_0 \triangleq \gamma_{\text{th}}^{-1} - N^{-1} \rho_0^{-1}$ ,  $\theta_2$  can be calculated as

$$\theta_2 = \mathbb{1}\{t_0 \geq 1\} + (1 - (1 - t_0)^N) \mathbb{1}\{0 \leq t_0 \leq 1\}. \quad (30)$$

Three approximations to  $\theta_r$  for  $r \geq 3$  and large  $N$  are described below. Approximating  $\rho_{\text{max}}^1$  as  $\rho_1^1$ , and assuming  $u_m$  as i.i.d. Gamma distributed with shape  $r-1$  and rate  $N$ , we obtain the Gamma approximation:

$$\text{Gamma: } \theta_r = 1 - \Gamma_{\text{inc}}(r-1, Nt_0) / \Gamma(r-1). \quad (31)$$

Approximating  $\rho_{\text{max}}^1 = \rho_1^1$  and  $u_m \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}((r-1)\mu_N, (r-1)\sigma_N^2)$ , where  $\mu_N \triangleq (N+1)^{-1}$ , and  $\sigma_N^2 \triangleq N(N+1)^{-2}(N+2)^{-1}$ , we obtain the Normal approximation:

$$\text{Normal: } \theta_r = 1 - \mathcal{Q}\left(\frac{t_0 - (r-1)\mu_N}{\sqrt{r-1}\sigma_N}\right), \quad (32)$$

where  $\mathcal{Q}(\cdot)$  is the standard Normal Q-function. Finally, in the Deterministic approximation, the SINR becomes  $\rho_m^k = N/(\rho_0^{-1} + r - k)$ , and  $\theta_r$  becomes

$$\text{Deterministic: } \theta_r = \mathbb{1}\{r \leq \lfloor N/\gamma_{\text{th}} - \rho_0^{-1} + 1 \rfloor\}. \quad (33)$$

*Proof.* See Appendix A. ■

*Remark:* The above approximations provide closed form expressions for  $\theta_r$  and are valid when  $N$  is large [23]. The first two approximations have SINRs that are obtained by applying the theory of deterministic equivalents to only the norms of the channels, and yields an SINR that is affected only by the randomness in the multi-user interference components. This is supported by the fact that the interference components converge to their deterministic equivalents slower than the norms converge to their deterministic equivalents [23]. The deterministic approximation follows directly from Lemma 1, where the SINR is a deterministic quantity, and hence  $\theta_r$  is a binary function of  $r$ . With finite number of antennas, due to small scale fading, the SINR of the users vary around this approximate SINR. These variations affect the value of  $\theta_r$ , and are not captured by the deterministic approximation, even though we obtain simple closed form expressions for it. As a consequence, the throughput computed using the deterministic

approximation can be far from the actual throughput in certain regimes and close to the actual throughput in other regimes, as will be seen in Sec. VI-A.

## VI. NUMERICAL RESULTS

In this section, the previously derived SINR analysis is used to evaluate the throughput of IRS with estimated channels via Monte Carlo simulations, and provide insights into the dependence of the system performance on the various system parameters. In each simulation, independent realizations of the user locations, the APM, and the fades experienced by the users are generated. The throughput for each simulation is calculated as described in Sec. II-A, and the effective system throughput  $\mathcal{T}$  is calculated by averaging over the simulations.

The results in this section are for  $T = 50$  RBs,  $N_s = 10^3$  Monte Carlo runs,  $\lambda = 10^{-2}$ ,  $\alpha = 3.76$ ,  $\sigma_h^2 = 1$ , SINR threshold  $\gamma_{\text{th}} = 10$ , MSBL threshold  $\gamma_{\text{pr}} = 10^{-6}$ , cell radius  $r_{\text{max}} = 1000$  m, and reference distance  $r_0 = 100$  m [20]. The number of users contending for the  $T$  RBs is computed based on the load  $L$  as  $M = \lfloor LT \rfloor$ . The soliton distribution [10] with  $d_{\text{max}} = 27$  maximum repetitions is used to generate the repetition factor  $d_m$  for the  $m$ th user, whose access pattern is formed by uniformly randomly choosing  $d_m$  RBs from  $T$  RBs [6]. The APM is formed by stacking the pattern vectors of all the users. The location of each user is uniformly sampled from within a cell of radius  $r_{\text{max}}$  centered at the BS. The path loss coefficient is calculated as  $\beta_m = (r_m/r_0)^{-\alpha}$  where  $r_m$  is the radial distance of the  $m$ th user from the BS. The signal to noise ratio (SNR) for the  $m$ th user is calculated as  $P\sigma_h^2\beta_m/N_0$ . The received SNR of a user at the edge of the cell at the BS is termed as the *cell edge SNR*, and is denoted by  $\text{SNR}_{\text{edge}}$ . The power levels of all users is chosen such that the signal from a user at a distance  $r_{\text{max}}$  from the BS is received at  $\text{SNR}_{\text{edge}}$ . This ensures that all users' signals are received at an SINR that at least  $\text{SNR}_{\text{edge}}$  on average, in *singleton* RBs. If  $\text{SNR}_{\text{edge}} \geq \gamma_{\text{th}}$ , i.e., it is such that the cell edge user's signal is decodable, then all users' signals are decodable with high probability in singleton RBs. The power levels of users is set to  $P = P^p = 20$  dBm [20] and  $N_0$  is chosen such that the cell edge SNR is 10 dB, unless otherwise stated.<sup>9</sup> The pilot sequence for each user is generated as  $\mathbf{p}_m \stackrel{\text{i.i.d.}}{\sim} \mathcal{CN}(\mathbf{0}_\tau, P^p \mathbf{I}_\tau)$ . The effect of different pilot sequences is studied in [16].

Fig. 2 shows the effect of pilot length on the system throughput at different  $L$  under the three estimation schemes, with  $N = 16$ . MMSE scheme performs the best and reaches the optimal throughputs of  $\mathcal{T} = L$  for very low pilot lengths. MSBL scheme achieves the optimal throughputs for  $L = 1, 2, 3$  at  $\tau = 4, 8, 12$  respectively, and beyond that, the performance is the same as that of MMSE. This shows that with a few additional pilot symbols, we can do away with the assumption of knowing the APM and path loss coefficients. LCMMSE scheme matches MMSE for  $L = 1$  and for higher  $L$ , it needs a lot more pilot symbols. This is because of both pilot contamination and low quality channel

<sup>9</sup>We consider equal pilot and data power for simplicity. Via simulations, we have observed that pilot power boosting can yield good improvement in the throughput, especially at cell edge SNRs close to 0 dB.

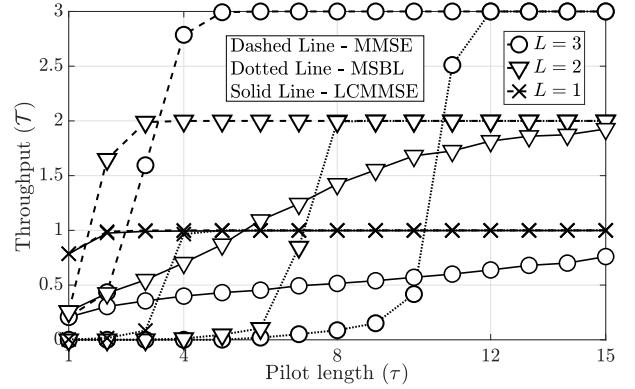


Fig. 2: Comparison between MMSE, MSBL, and LCMMSE schemes.

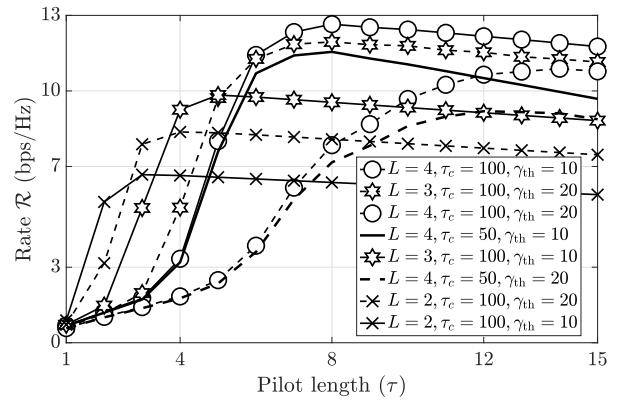


Fig. 3: Impact of pilot length  $\tau$  on rate with MMSE.

estimates. Also, we note that the use of orthogonal pilots would require  $\tau \geq 50, 100, 150$  for  $L = 1, 2, 3$ , respectively. The optimal throughput of  $\mathcal{T} = L$  is achieved with far fewer pilot symbols under all the three estimation schemes. This is because a small subset of users transmit in any RB in IRS. Finally, under all the three schemes, we can achieve  $\mathcal{T} \geq 1$ , which is the maximum throughput achievable under perfectly coordinated orthogonal access, i.e., grant-based orthogonal access. This shows the utility of using IRS as a GFRA protocol for mMTC, especially due to its high performance at medium to high  $L$ . To summarize, the pilot length has a significant impact on the performance of IRS and yields near-optimal throughputs at significantly lower pilot lengths than that required for orthogonal pilot transmission. The drop in  $\mathcal{T}$  at low pilot lengths under estimated channels underscores the importance of accounting for the effect of imperfect CSI in analyzing the performance of IRS.

We focus on MMSE/LCMMSE hereafter in order to avoid clutter in the plots, since MSBL matches the performance of MMSE with slightly higher  $\tau$ . In Fig. 3, we investigate the effect of  $L$ ,  $\tau_c$  and  $\gamma_{\text{th}}$  on the achievable rate  $\mathcal{R}$  of the system with MMSE, with  $N = 16$ . Here, the rate is obtained as  $\mathcal{R} = (1 - \tau/\tau_c)\mathcal{T} \log_2(1 + \gamma_{\text{th}})$  (bps/Hz), where  $\tau_c$  is the total length of any user's packet. Firstly, we look at the effect of changing  $\gamma_{\text{th}}$  by fixing  $\tau_c = 100$ . For  $L = 2$ ,  $\gamma_{\text{th}} = 20$  offers a higher rate than  $\gamma_{\text{th}} = 10$ , provided  $\tau \geq 3$ . Thus, at low

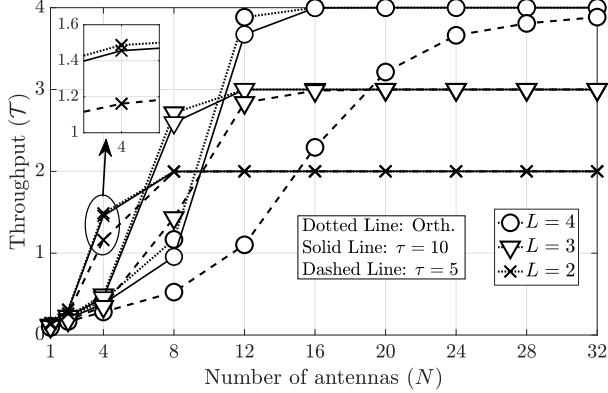
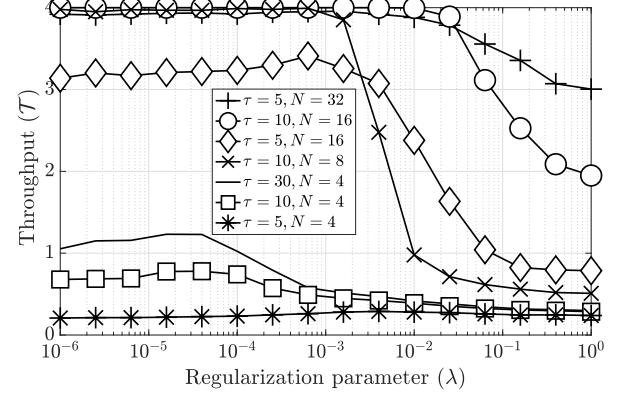
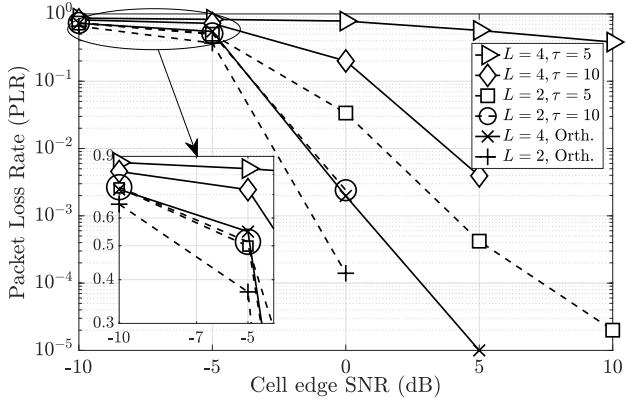
Fig. 4: Effect of number of antennas  $N$  with MMSE.Fig. 6: Effect of regularization parameter and  $\tau$  with MMSE.

Fig. 5: Effect of cell edge SNR with MMSE.

loads, increasing  $\gamma_{\text{th}}$  (correspondingly, selecting a higher order modulation and coding scheme) leads to better achievable rates. In contrast, when  $L = 4$ ,  $\gamma_{\text{th}} = 10$  outperforms  $\gamma_{\text{th}} = 20$ , because the system is highly interference limited. Next, comparing  $L = 2, 3, 4$  for  $\tau_c = 100$  and  $\gamma_{\text{th}} = 10$ , we see that the rate improves with  $L$ , provided the pilot length is large enough. Finally, decreasing  $\tau_c$  reduces the achievable rate, as the relative overhead due to pilots increases. Thus, at high loads, the throughput  $\mathcal{T}$  limits the achievable rate, while at low loads, the SINR threshold  $\gamma_{\text{th}}$  is the primary factor in determining the achievable rate.

In Fig. 4, we investigate the effect of the number of antennas at the BS, by plotting the throughput with MMSE channel estimation for different  $L$  and  $\tau$ . Intuitively, we expect that, to achieve the optimal throughput of  $L$ , we would require slightly more than  $L\bar{d}$  antennas at the BS, since  $L\bar{d}$  users transmit packets per RB on average. The orthogonal pilots curve is obtained by allocating  $\tau = M = \lfloor LT \rfloor$  for each  $L$ . Under all configurations, it is observed that increasing  $N$  has a significant impact, and the peak throughput achieved reaches its maximum of  $\mathcal{T} = L$ . Further,  $\tau = 10$  achieves a very similar performance as that of orthogonal pilots, and  $\tau = 5$  performs poorly at low  $N$  and high  $L$ . For  $L = 2$ , the throughput reaches the peak  $\mathcal{T} = 2$  for  $N \geq 8$  for all three values of  $\tau$ . Similarly, for a high load of  $L = 3$ , the throughput reaches the peak,  $\mathcal{T} = 3$ , for  $N \geq 16$ . For  $L =$

2, 3, since the average repetition factor  $\bar{d} = 3$ , each RB is occupied by 6, 9 users, respectively. Thus, a slightly higher number of antennas is sufficient to recover all the packets, provided accurate channel estimates are available (i.e.,  $\tau$  is large enough). It is observed that at  $L = 2$ ,  $N = 4$  and  $L = 3$ ,  $N = 8$ , improving  $\tau$  greatly improves the throughput. Increasing the number of antennas increases the array gain and the decoding capability of the regularized zero forcing decoder at the BS, which in turn leads to more users getting decoded. This shows the effectiveness of the number of antennas in improving the throughput. Also, when  $N = 12$ , the dramatic drop in the throughput of  $\mathcal{T} = 3.8$  for  $\tau = 200$  (orthogonal pilots) to  $\mathcal{T} = 1.2$  for  $\tau = 5$ , which is around 70% loss in performance, shows that it is crucial to account for estimated CSI while analyzing the performance of IRSAs.

Fig. 5 shows the impact of cell edge SNR on the packet loss rate PLR with MMSE, with  $N = 16$ . For  $\text{SNR} < -5$  dB, the PLR is high, and in the noise-limited regime ( $-5 < \text{SNR} < 0$  dB), an increase in cell edge SNR sharply decreases the PLR. For  $L = 4$ ,  $\tau = 5$ , the system becomes interference-limited, and thus the performance saturates at high SNR. This is because, at low  $\tau$ , both signal and interference powers get scaled equally, and the SINR remains roughly constant. Increasing  $\tau$  from 5 to 10 and then to orthogonal pilots, we observe that the PLR falls from  $0.5$  to  $10^{-2.5}$  to  $10^{-5}$ . The higher  $\tau$  and SNR result in accurate channel estimates, and thus very low PLR is observed. Similarly, at  $L = 2$ , the drop of PLR from  $10^{-1.7}$  to  $10^{-2.8}$  to  $10^{-3.9}$  for  $\tau = 5, 10$  and orthogonal pilots emphasizes the need to account for estimated CSI when analyzing the performance of IRSAs. In summary, the overall performance can be improved by increasing the pilot length, number of antennas, or cell edge SNR, but these need to be increased judiciously, keeping the other parameters in mind.

Fig. 6 shows the effect of the regularization parameter,  $\lambda$ , on the throughput of the system when MMSE estimation is employed, with  $L = 4$ . As  $\lambda$  is varied from  $10^{-6}$  to 1, the curves go from ZF on the left to RZF in the middle and finally to MRC on the right. For  $N = 4$ , increasing  $\tau$  from 5 to 10 to 30 only marginally improves the throughput. This is because the system is highly interference limited, and hence

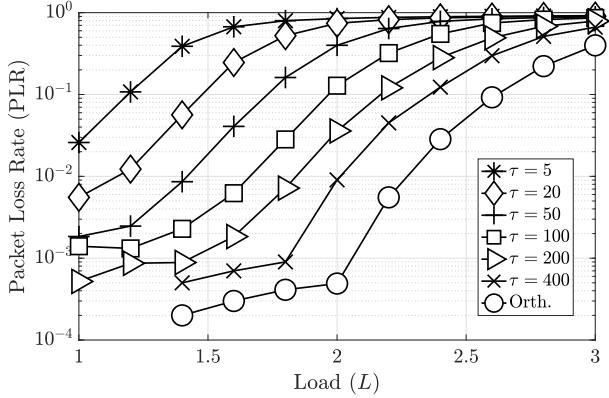


Fig. 7: Impact of load on PLR with LCMMSE.

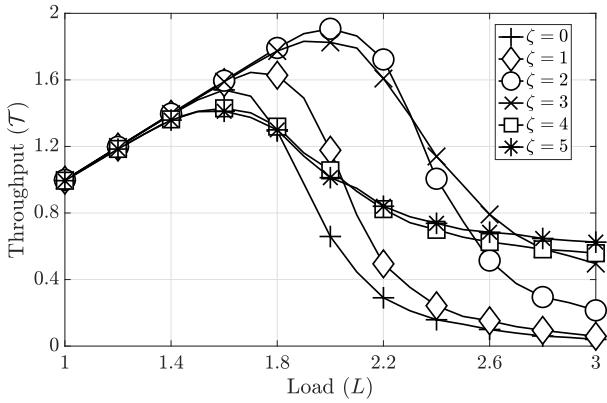
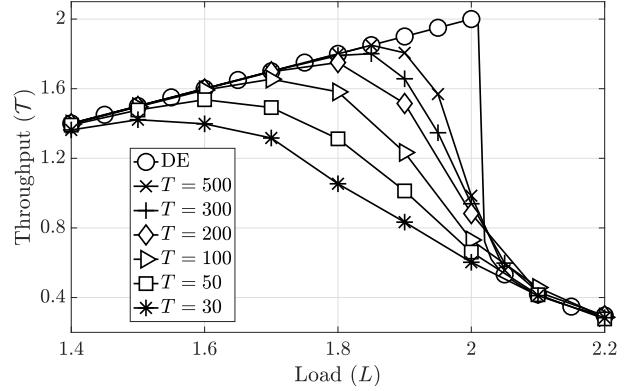


Fig. 8: Impact of power control on throughput.

channel inversion does not work well at low  $N$ . For  $\tau = 5$ , increasing  $N$  from 4 to 16 to 32 improves the performance due to the interference suppression capability of RZF. Similar observations can be made for  $\tau = 10$  as well. MRC does not have the interference suppression capability of RZF, and thus the performance saturates at a low value for all  $\tau$ . We note that the optimal throughput of  $T = 4$  is obtained over a wide range of  $\lambda$ , and thus precise optimization of  $\lambda$  is not necessary to obtain near-optimal throughputs.

Fig. 7 studies the impact of  $L$  and  $\tau$  on the system packet loss rate, PLR, evaluated with  $N = 16$ ,  $\gamma_{\text{th}} = 16$ , and  $\lambda = 1$ . As the pilot length  $\tau$  increases, better quality channel estimates are obtained, and the corresponding SINR increases. In particular, the system requires higher pilot lengths due to the use of LCMMSE estimates. The loss rates reduce with increase in  $\tau$ , and gets closer to the orthogonal loss rate. The PLR of perfectly coordinated orthogonal access is the lowest. Similar to existing works, there is an error floor region where the PLR is very low (upto  $L = 2$  for orthogonal pilots) after which the PLR increases rapidly and is called the waterfall region. Here  $L = 2$  marks the inflection load, where the system transitions from the error floor to the waterfall region.

In Fig. 8, the impact of power control on the throughput with LCMMSE is characterized. For this plot, users transmit at powers that are dependent on their distances from the BS. Specifically, the  $m$ th user, who is located at a distance  $r_m$

Fig. 9: Effect of  $T$  on the throughput.

from the BS transmits at a power  $P(r_m/r_0)^{\alpha-\zeta}$ , making  $\zeta$  the effective path loss exponent. The cell edge SNR is fixed to 10 dB, and the throughputs are obtained by varying  $\zeta$  and  $P$ . When  $\zeta = 0$ , the signals of the users undergo pure fading, and the system achieves a peak throughput of  $T = 1.52$  at  $L = 1.6$ . Further, as  $L$  is increased, the throughput drops to 0. The throughput of the system increases as  $\zeta$  increases, until  $\zeta = 2/3$ . The exact  $\zeta$  that yields the highest throughput is dependent on other system parameters such as SNR,  $\gamma_{\text{th}}$ , and  $N$ . As  $\zeta$  is increased, the channel coefficients of the users become more disparate, and thus offer a higher degree of capture effect. Beyond  $\zeta = 3$ , the throughput decreases as the exponent is so high that the received signal power becomes comparable to the noise. For higher  $\zeta$ , the throughput saturates as  $L$  is increased since a few users are always decoded due to path loss disparity. The channel fades and the path loss coefficients contribute to the disparity amongst the channel coefficients of the users, and thus such a system has higher throughputs than a system with only path loss [9] or only fading [11]. Thus, it is useful to consider the combined effects of fading and path loss in optimizing the performance.

#### A. Theoretical Validation of Throughput

The results in this subsection are presented for  $\tau = 10$ , cell edge SNR = 10 dB,  $N = 16$ ,  $\lambda = 10^{-2}$ ,  $\gamma_{\text{th}} = 16$ ,  $d_{\text{max}} = 8$  maximum repetitions and  $N_s = 10^3$  Monte Carlo runs. To reduce clutter in the plots, we present the results for the lowest complexity (LCMMSE) channel estimation scheme.

Fig. 9 investigates the effect of increasing the number of RBs on the throughput. The peak throughput increases from  $T = 1.52$  at  $L = 1.6$  for  $T = 50$  to  $T = 1.85$  at  $L = 1.85$  for  $T = 500$ . Since  $\bar{d}$  is fixed, each user has a larger number of RBs to choose from as  $T$  is increased. Thus, the interference reduces, and the throughput increases until it reaches a peak and then drops off. The success probability  $\theta_r$  is evaluated empirically via  $10^4$  Monte Carlo runs, and this in turn yields the asymptotic theoretical throughput, which is marked as "DE". This can be achieved as  $M, T \rightarrow \infty$  with a fixed  $L$ . It is seen that this asymptotic throughput increases linearly with the load until it hits a maximum at the inflection load of the system, which occurs at  $L^* = 2$  in this case. The throughput

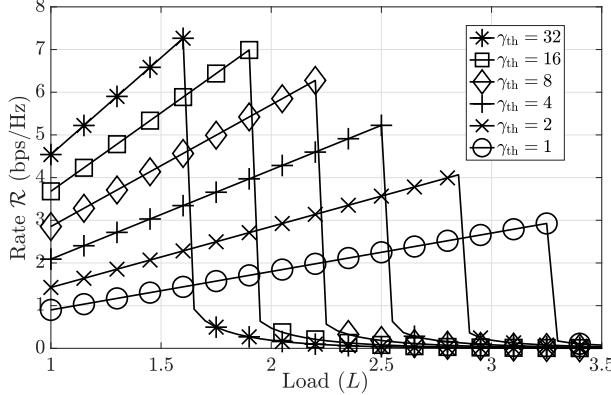


Fig. 10: Rate for different SINR thresholds.

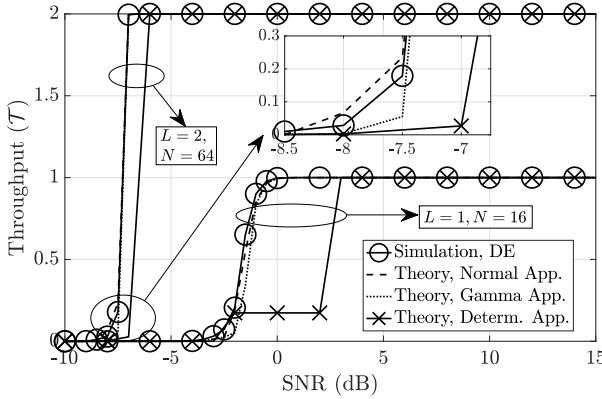


Fig. 11: Validation of theoretical approximations.

drops sharply beyond this load. The asymptotic throughput provides an upper bound on the throughput achievable with finitely many RBs for low to moderate loads. At very low and high loads, the throughput achieved with finitely many RBs exactly matches with the DE asymptotic throughput. A convenient operating point would be to set the system load to, say, 90% of the inflection load, as, in this case, only finitely many RBs would be sufficient to achieve the asymptotic throughput. Finally, it can be observed that the throughput of the system can be increased by increasing  $T$ , but only when the system is operated at a load that is lower than the inflection load. Beyond the inflection load, the system is always interference-limited and increasing  $T$  does not help.

In Fig. 10, the asymptotic rate of the system is plotted versus the system load for different SINR thresholds with  $\tau_c = 100$ . For a fixed  $\gamma_{\text{th}}$ ,  $\mathcal{R}$  increases until the inflection load and then drops off to zero. It is observed that a high  $\mathcal{R}$  can be achieved at lower loads by choosing a high  $\gamma_{\text{th}}$ , whereas, at high loads, in order to serve more users,  $\gamma_{\text{th}}$  must be kept low. The choice of the threshold  $\gamma_{\text{th}}$  decides the rate of transmission, which in turn is related to the modulation and coding scheme to be used. In summary, the SINR threshold  $\gamma_{\text{th}}$ , which depends on the modulation and coding scheme employed and determines the data rate, can be chosen based on the system parameters such as the number of antennas, training duration, number of users/RBs, and the transmit power.

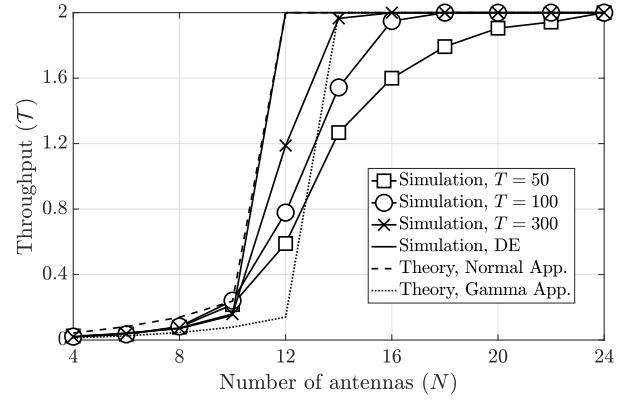


Fig. 12: Comparison of approximations with simulation.

We now validate the approximations derived in Theorem 3 with the simulations obtained with MRC,  $d_{\max} = 27$  maximum repetitions, and  $\gamma_{\text{th}} = 10$ . Fig. 11, reveals an inflection SNR\* of 0 dB and -7 dB for  $L = 1, N = 16$  and  $L = 2, N = 64$  respectively, which behaves similar to the inflection load  $L^*$ . Both the normal and the gamma approximations match well with the asymptotic throughput obtained from the DE process. This is because the deterministic approximation results in an SINR that is completely deterministic and  $\theta_r$  that is a binary function of  $r$ , and consequently does not capture the statistics of the SINRs very well. Further, the deterministic approximation results in a throughput that acts as a step function since  $\theta_r$  depends binarily on  $N, \gamma_{\text{th}}$ , and SNR. As we go from  $L = 1, N = 16$  to  $L = 2, N = 64$ , the approximations become closer, and both the normal and the gamma approximations match perfectly with the asymptotic throughput. In summary, the theoretical curves with the approximations match the simulations when  $N$  is increased, as expected.

Fig. 12 examines the effect of  $T$  on the approximations with  $L = 2$  and SNR = 10 dB. With finitely many RBs, such as  $T = 50, 100, 300$ , the throughput achieves the optimal throughput  $T = 2$  for  $N = 24, 18, 16$ . The asymptotic throughput obtained with DE provide an inflection  $N^* = 12$ , which matches perfectly with the normal approximation. The gamma approximation does not match as well as the normal approximation. Here, the curves are with MRC and perfect CSI, and the presented curves are valid upper bounds to the throughputs with estimated CSI. These can be achieved with high enough  $\tau$  as observed in Fig. 2, and thus the derived results provide very good approximations to the asymptotic throughput achievable with estimated CSI.

## VII. CONCLUSIONS

This paper studied the effect of estimated CSI on the throughput of IRSAs, which is a distributed medium access protocol for mMTC involving repetition of packets across different randomly selected RBs. Decoding the users' packets at the BS involves successive interference cancellation. First, the channel estimates were derived under three schemes: a sparsity-based scheme with MSBL, MMSE, and LCMMSE.

The corresponding SINR of all the users were obtained under all three schemes accounting for pilot contamination, channel estimation errors, path loss as well as multiple antennas at the BS. It was seen that these errors significantly reduce the peak achievable throughput, even resulting in up to 70% loss in certain regimes. Further, a density evolution based analysis was presented to characterize the asymptotic performance of the protocol when users perform path loss inversion based power control. Here, several approximations to the success probability  $\theta_r$  were derived and it was seen that these approximations match well as the number of antennas at the BS becomes large. Finally, several new insights into the design of IRS-based systems was discussed, namely, the improvement of the system throughput, the evaluation of the operating load beyond which the system becomes interference limited, and the choice of the decoding threshold  $\gamma_{\text{th}}$ . The results underscored the importance of accounting for practical channel estimation in studying the throughput offered by the IRS protocol. Future work could involve using differential evolution techniques [24] to obtain the optimal repetition distribution that maximizes the throughput in the finite frame length regime.

#### APPENDIX A: PROOF OF THEOREM 3

Let  $k$  denote the intra-RB decoding iteration. When perfect CSI is available at the BS and the users perform path loss inversion, the SINR of the  $m$ th user in an RB is computed as

$$\rho_m^k = \frac{P\|\mathbf{h}_m\|^4}{N_0\|\mathbf{h}_m\|^2 + P\sum_{i \in \mathcal{S}_k^m} |\mathbf{h}_m^H \mathbf{h}_i|^2}. \quad (34)$$

For  $r = 1$ ,  $\rho_1^1 = P\|\mathbf{h}_m\|^2/N_0$ , and  $\theta_1$  reduces to

$$\theta_1 = \Pr(\rho_1^1 \geq \gamma_{\text{th}}) = \Gamma_{\text{inc}}(N, \rho_0^{-1}\gamma_{\text{th}})/\Gamma(N), \quad (35)$$

where  $\rho_0 \triangleq P\sigma_h^2/N_0$ ,  $\Gamma_{\text{inc}}(s, x) = \int_x^\infty t^{s-1} e^{-t} dt$  is the upper incomplete gamma function and  $\Gamma(s)$  is the ordinary gamma function. The interference is written as  $t_{mi} = |\mathbf{h}_m^H \mathbf{h}_i|^2/(\|\mathbf{h}_m\|^2\|\mathbf{h}_i\|^2)$ , where  $t_{mi} \sim \text{Beta}(\alpha = 1, \beta = N)$ . We use  $\xrightarrow{\text{a.s.}}$  to denote convergence in the almost surely sense. Since  $\|\mathbf{h}_i\|^2/N \xrightarrow{\text{a.s.}} \sigma_h^2$  and  $\|\mathbf{h}_i\|^4/N^2 \xrightarrow{\text{a.s.}} \sigma_h^4$  as  $N \rightarrow \infty$  [23], we can approximate the SINR as

$$\rho_m^k \approx N(\rho_0^{-1} + N\sum_{i \in \mathcal{S}_k^m} t_{mi})^{-1}. \quad (36)$$

Here, we have applied the theory of deterministic equivalents to only the channel norms and not to the interference. This is supported by the fact that the interference components converge to their deterministic equivalents slower than the norms converge to their deterministic equivalents [23].

For  $r = 2$ , since  $t_{12} = t_{21}$ ,  $\rho_1^1 = \rho_2^1 = N/(\rho_0^{-1} + Nt_{12})$ . Thus,  $\rho_{\max}^1 = N/(\rho_0^{-1} + Nt_{12})$  and  $\rho_{\max}^2 = N\rho_0$  with  $\rho_{\max}^1 \leq \rho_{\max}^2$ . Thus, the success probability reduces to  $\theta_r = \Pr(\rho_{\max}^1 \geq \gamma_{\text{th}})$ . Let  $t_0 \triangleq \gamma_{\text{th}}^{-1} - N^{-1}\rho_0^{-1}$ . Hence,  $\theta_2$  is calculated as

$$\theta_2 \approx \Pr(\rho_{\max}^1 \geq \gamma_{\text{th}}) = \Pr(t_{12} \leq t_0) = \mathbb{1}\{t_0 \geq 1\} + (1 - (1 - t_0)^N)\mathbb{1}\{0 \leq t_0 \leq 1\}. \quad (37)$$

For  $r \geq 3$ ,  $\rho_m^k$  need not be a monotonically increasing function of  $k$  as seen in (36), and thus we cannot order the SINRs

meaningfully to compute a closed form expression for  $\theta_r$ . With  $u_m = \sum_{i \in [r] \setminus m} t_{mi}$ , the maximum SINR in the first intra-RB iteration is calculated as  $\rho_{\max}^1 = \max_{m \in [r]} N(\rho_0^{-1} + Nu_m)^{-1}$ . Here,  $u_m$  is not independent across  $m$  and it is not clear which  $u_m$  is the minimum. Thus, we approximate  $\rho_{\max}^1$  as  $\rho_1^1$ , and upon dropping the other SINR terms from (28),  $\theta_r$  becomes

$$\theta_r \approx \Pr(\rho_1^1 \geq \gamma_{\text{th}}) = \Pr(u_m \leq t_0). \quad (38)$$

We now discuss two approximations to  $u_m$  to evaluate  $\theta_r$ , with the assumption that  $u_m$  is independent across  $m$ .

Since  $\lim_{N \rightarrow \infty} \text{Beta}(\alpha = 1, \beta = N) = \exp(\lambda = N)$ , we approximate  $t_{mi} \sim \exp(N)$ , which is a good approximation at high  $N$  [22]. Even with this approximation,  $u_m$  is identically Gamma distributed across users but not independent. Thus, with the independence assumption,  $u_m$  is i.i.d. Gamma distributed with shape parameter  $r - 1$  and rate parameter  $N$ , i.e.,  $u_m \xrightarrow{\text{i.i.d.}} \text{Gamma}(r - 1, N)$ . Thus, we obtain the *Gamma* approximation:

$$\theta_r \approx 1 - \Gamma_{\text{inc}}(r - 1, Nt_0)/\Gamma(r - 1). \quad (39)$$

Similarly, when we assume  $t_{mi}$  is Normal distributed,  $u_m$  is identically Normal distributed across users but not independent. Let  $\mu_N = (N + 1)^{-1}$  and  $\sigma_N^2 = N(N + 1)^{-2}(N + 2)^{-1}$ . If we approximate  $t_{mi} \sim \mathcal{N}(\mu_N, \sigma_N^2)$  and  $u_m$  is independent across  $m$ , then  $u_m \xrightarrow{\text{i.i.d.}} \mathcal{N}((r - 1)\mu_N, (r - 1)\sigma_N^2)$ . Thus, we obtain the *Normal* approximation:

$$\theta_r \approx 1 - \mathcal{Q}\left(\frac{t_0 - (r - 1)\mu_N}{\sqrt{r - 1}\sigma_N}\right), \quad (40)$$

where  $\mathcal{Q}(\cdot)$  is the standard Normal Q-function.

A simpler expression can be obtained for  $\theta_r$  by applying the theory of deterministic equivalents to not just the channel norms but also to the interference. Thus,  $|\mathbf{h}_i^H \mathbf{h}_m|^2/N \xrightarrow{\text{a.s.}} \sigma_h^4$ , as  $N \rightarrow \infty$  [23]. Thus, the SINR becomes

$$\rho_m^k = N/(\rho_0^{-1} + r - k), \quad (41)$$

which is not random and is a deterministic function of  $N$  and  $\rho_0$ . This expression for SINR follows from Lemma 1. Thus, we obtain the *deterministic* approximation:

$$\theta_r = \Pr(\rho_1^1 \geq \gamma_{\text{th}}) = \mathbb{1}\{r \leq \lfloor N/\gamma_{\text{th}} - \rho_0^{-1} + 1 \rfloor\}. \quad (42)$$

#### REFERENCES

- [1] C. R. Srivatsa and C. R. Murthy, "Throughput analysis of PDMA/IRSA under practical channel estimation," in *Proc. SPAWC*, July 2019.
- [2] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, 2015.
- [3] T. Xia, M. M. Wang, C. Jiang, J. Zhang, L. Wang, and X. You, "Practical machine-type communication for energy internet of things: An introduction," *IEEE Commun. Mag.*, vol. 3, no. 1, pp. 48–59, 2019.
- [4] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the internet of things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, 2018.
- [5] L. Liu and W. Yu, "Massive connectivity with massive MIMO—part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, 2018.
- [6] G. Liva, "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 477–487, February 2011.

- [7] E. Paolini, G. Liva, and M. Chiani, "Coded slotted ALOHA: A graph-based method for uncoordinated multiple access," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6815–6832, Dec 2015.
- [8] S. Verdú *et al.*, *Multiuser detection*. Cambridge university press, 1998.
- [9] E. E. Khaleghi, C. Adjih, A. Alouini, and P. Muhlethaler, "Near-far effect on coded slotted ALOHA," in *Proc. PIMRC*, Oct 2017.
- [10] K. R. Narayanan and H. D. Pfister, "Iterative collision resolution for slotted ALOHA: An optimal uncoordinated transmission policy," in *Proc. ISTC*, Aug 2012, pp. 136–139.
- [11] F. Clazzer, E. Paolini, I. Mambelli, and C. Stefanovic, "Irregular repetition slotted ALOHA over the Rayleigh block fading channel with capture," in *Proc. ICC*, May 2017.
- [12] M. Ghanbarinejad and C. Schlegel, "Irregular repetition slotted ALOHA with multiuser detection," in *Proc. WONS*, 2013, pp. 201–205.
- [13] J. Su, G. Ren, B. Zhao, and J. Ding, "Enhancing irregular repetition slotted ALOHA with polarization diversity in LEO satellite networks," *KSII TIIS*, vol. 14, no. 9, pp. 3907–3923, 2020.
- [14] F. Clazzer, C. Kissling, and M. Marchese, "Enhancing contention resolution ALOHA using combining techniques," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2576–2587, 2017.
- [15] F. Clazzer, F. Lázaro, G. Liva, and M. Marchese, "Detection and combining techniques for asynchronous random access with time diversity," in *SCC 2017*, 2017, pp. 1–6.
- [16] C. R. Srivatsa and C. R. Murthy, "User activity detection for irregular repetition slotted ALOHA based MMTC," *Accepted, IEEE Trans. Signal Process.*, June 2022.
- [17] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access-a novel nonorthogonal multiple access for fifth-generation radio networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3185–3196, April 2017.
- [18] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, 2003.
- [19] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, 2007.
- [20] E. Björnson, J. Hoydis, L. Sanguinetti *et al.*, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Foundations and Trends® in Signal Processing*, vol. 11, no. 3-4, pp. 154–655, 2017.
- [21] Wang Lin, Xiao Juan, and Guanrong Chen, "Density evolution method and threshold decision for irregular LDPC codes," in *Proc. ICCCAS*, vol. 1, 2004, pp. 25–28 Vol.1.
- [22] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.
- [23] R. Couillet and M. Debbah, *Random matrix methods for wireless communications*. Cambridge University Press, 2011.
- [24] R. Storn and K. Price, "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, 1997.



**Chirag Ramesh Srivatsa** (S'20) received the B.E. degree in Electronics and Communications Engineering from PES Institute of Technology, Bangalore, India, in 2017. He is currently pursuing the Ph.D. degree in the Department of Cyber-Physical Systems, Indian Institute of Science, Bangalore, India. His research interests include signal processing for massive random access and massive MIMO communications.



**Chandra R. Murthy** (S'03–M'06–SM'11) received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology Madras, India, in 1998, the M.S. and Ph.D. degrees in Electrical and Computer Engineering from Purdue University and the University of California, San Diego, USA, in 2000 and 2006, respectively. In Sept. 2007, he joined the Department of Electrical Communication Engineering at the Indian Institute of Science, Bangalore, India, where he is currently working as a Professor.

His research interests are in the areas of energy harvesting communications, multiuser MIMO systems, and sparse signal recovery techniques applied to wireless communications. His paper won the best paper award in the Communications Track at NCC 2014 and a paper co-authored with his students won the student best paper award at the IEEE ICASSP 2018 and IEEE ISIT 2021. He has over 75 journal papers and over 100 conference papers to his credit. He was an associate editor for the IEEE SIGNAL PROCESSING LETTERS during 2012–16, the IEEE TRANSACTIONS ON COMMUNICATIONS during 2017–22, and the SADHANA JOURNAL during 2016–18. He was an elected member of the IEEE SPCOM Technical Committee for the years 2014–19. He is currently serving as an area editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and as an associate editor for the IEEE TRANSACTIONS ON INFORMATION THEORY.