Channel State Information Based User Censoring in Irregular Repetition Slotted Aloha

Chirag Ramesh Srivatsa, *Graduate Student Member, IEEE,* and Chandra R. Murthy, *Fellow, IEEE*
Dept. of CPS and Dept. of ECE, Indian Institute of Science, Bengaluru, India (e-mail: {chiragramesh, cmurthy}@iisc.ac.in).

ICC 2023 - IEEE International Conference on Communications | 978-1-5386-7462-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICC45041.2023.10278965

***Abstract*—Irregular repetition slotted aloha (IRSA) is a massive random access protocol which can be used to serve a large number of users while achieving a packet loss rate (PLR) close to zero. However, if the number of users is too high, then the system is interference limited and the PLR is close to one. In this paper, we propose a variant of IRSA in the interference limited regime, namely Censored-IRSA (C-IRSA), wherein users with poor channel states censor themselves from transmitting their packets. We theoretically analyze the throughput performance of C-IRSA via density evolution. Using this, we derive closed-form expressions for the optimal choice of the censor threshold which maximizes the throughput while achieving zero PLR among uncensored users. Through extensive numerical simulations, we show that C-IRSA can achieve a $4\times$ improvement in the peak throughput compared to conventional IRSA.**

***Index Terms*—Irregular repetition slotted aloha, massive machine-type communications, user censoring, random access**

I. Introduction

Massive machine-type communications (mMTC) is an evolving use-case, expected to serve about a million users per square km [1]. In this context, irregular repetition slotted aloha (IRSA) [2] is a distributed massive random access protocol which has received much attention in the literature [3], [4]. The performance of IRSA mainly depends on the system load, i.e., the number of users participating in the protocol per frame. At low system loads, the system is not interference-limited, and the packet loss rate (PLR) is close to zero. As the system load increases beyond a so-called *inflection load*, IRSA becomes interference limited, and the system throughput rapidly drops to zero, with PLR approaching one [3]. In this paper, we address the issue of the poor throughput of IRSA in the high load regime by introducing a *distributed self-censoring scheme,* which allows the system to maintain the throughput at the maximum possible value even as the load increases.

In IRSA, users transmit packet replicas in multiple randomly selected slots (within a frame) to a base station (BS); the latter decodes the packets using successive interference cancellation (SIC) [5]. If the BS successfully decodes a user in a slot, it uses the decoded data to perform SIC in all other slots in which that user has transmitted a replica. The decodability of any user in IRSA depends on the signal to interference plus noise ratio (SINR) of that user [6]. If the users have poor channel states or there are many collisions in a slot resulting in increased multi-user interference (MUI), then the SINR decreases, leading to users not getting decoded.

When the system load is low, all users with sufficiently good channel states can be decoded at the end of the SIC process [7]. However, as the load increases, more collisions lead to the system becoming interference limited, and the throughput quickly drops close to zero [5]. To tackle this issue, in this paper, we propose a modified IRSA protocol, as follows. The BS transmits a pilot signal at the start of each frame, using which the users estimate their channel state information (CSI). Then, users with poor CSI *self-censor*, i.e., they refrain from transmitting, thereby reducing collisions and enable the decoding of the transmissions from the active users to succeed. Here, choosing too high a (CSI-based) censor threshold leads to very few users participating in the IRSA protocol, while too low a threshold leads to too many collisions and the system becoming interference-limited. Thus, given the system load, there is an optimal threshold that maximizes the throughput. The censor threshold is calculated by the BS based on the load and the SNR, and its value is periodically broadcast to the users. Note that this approach retains the fully distributed nature of IRSA.

IRSA has been studied for the collision channel [2], with multiple antennas [3], with activity detection [4], with path loss [6], for the Rayleigh fading channel [7], with channel estimation errors [8], and with multi-cell effects [9]. Density evolution has been used to characterize the asymptotic throughput of IRSA [3], [6], [7]. Variants of aloha such as K-repetition have also been studied [10], [11]. However, none of these papers address the dramatic increase in PLR and the corresponding reduction in throughput as the system load increases, which is our focus in this paper. Our specific contributions are as follows:

1) We propose a censored-IRSA (C-IRSA) protocol in the interference limited regime, where users with poor CSI self-censor to decrease the effective system load, thereby enabling the uncensored users to be decoded at the BS.
2) We theoretically analyze the asymptotic performance of C-IRSA using density evolution.
3) We provide the optimal choice of the censor threshold with which the PLR of uncensored users can be driven close to zero at all system loads, while maintaining the throughput of the system at its highest value.

With CSI-based censoring in C-IRSA, we can achieve a $4\times$ improvement in the peak performance of the system compared to conventional IRSA. Further, C-IRSA can be operated at the peak performance for all system loads, whereas the throughput of conventional IRSA becomes zero at high loads.

Notation: The symbols a, $\mathbf{a}$, and $\mathbf{A}$, denote a scalar, a vector, and a matrix, respectively. $[N]$ denotes the set $\{1, 2, \ldots, N\}$. $\mathbb{1}\{\cdot\}$, $|\cdot|$, $[\cdot]^*$, and $\mathbb{E}[\cdot]$, denote the indicator, magnitude (or cardinality of a set), conjugate, and expectation, respectively.

978-1-5386-7462-8/23/$31.00 ©2023 IEEE

II. System Model

We consider an IRSA system with M single-antenna users communicating with a central BS over a frame consisting of T slots. The BS is located at the cell center, and the users are arbitrarily located within the cell. mMTC applications use similar settings as narrowband internet of things, which uses a narrow bandwidth of 180 kHz [1]. Over this band, the channel can be assumed to be flat and Rayleigh block fading. The BS allocates a pre-specified band to all the users in the system and the M users transmit their packets within this band. The *system load,* L, is defined as the ratio of the number of users to the number of slots, $L \triangleq M/T$. In conventional IRSA, in each frame, the users randomly select a subset of the slots, and transmit replicas of their packets in the chosen slots. The access of the T slots in a given frame can be represented as a *binary access pattern matrix* $\mathbf{G} \in \{0,1\}^{T\times M}$ [3]. The (t,m)th element of $\mathbf{G}$, denoted by g_{tm}, equals 1 if the mth user transmits its packet in the tth slot, and $g_{tm} = 0$ otherwise. In such a protocol, when L is high, there will be too many collisions in any slot, leading to a failure of the SIC-based decoding process (described below), resulting in low throughput.

The C-IRSA protocol we propose works as follows. At the start of each frame, the BS transmits a pilot signal, using which the users estimate their channel state. The users participate in the IRSA protocol if and only if the magnitude squared of their channel exceeds a censor threshold denoted by ν. We refer to the users who self-censor as *inactive* or *censored* users, and the other users as *active* or *uncensored* users. A censored user can sleep till the next time it has data to transmit, by when its channel state would change. At the BS, the active users' packets are decoded using the SIC process as with the conventional IRSA protocol.

The mth user transmits a symbol x_m with $\mathbb{E}[x_m]=0$ and $\mathbb{E}[|x_m|^2]=1$. The received signal y_t at the BS in the tth slot is

$$y_t = \textstyle\sum_{m=1}^{M} \sqrt{\rho_0} a_m g_{tm} h_m x_m + n_t, \tag{1}$$

where $h_m \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0,1)\ \forall m \in [M]$ is the uplink fading channel of the mth user, assumed independent across users and frames; $a_m = \mathbb{1}\{|h_m|^2 \geq \nu\}$ is the activity coefficient of the mth user, and $n_t \overset{\text{i.i.d.}}{\sim} \mathcal{CN}(0,1)$ is the complex AWGN at the BS. Also, $\rho_0 \triangleq P\sigma_{\mathtt{h}}^2/N_0$ denotes the signal to noise ratio (SNR) of any user (in the absence of any collisions), where P is the transmit power, $\sigma_{\mathtt{h}}^2$ is the fading variance, and N_0 is the noise variance. Here, we assume that the users perform path loss inversion based power control to ensure the same average received power levels of all users, which in turn ensures fairness. In addition, inverting only the path loss rather than full channel inversion helps with *capture effect*, which allows some of the users to be decoded in slots where there are collisions, improving the throughput [3]. We denote the set of active users by $\mathcal{A} \triangleq \{i \in [M] | |h_i|^2 \geq \nu\}$, the number of active users by $M_a \triangleq |\mathcal{A}|$, and the *active load* L_a by $L_a \triangleq M_a/T$.

1) SIC-based Decoding: The BS iteratively processes the received signal. In each slot, the BS attempts to decode the users' packets. If a user is successfully decoded, which can be verified via a cyclic redundancy check, then using the decoded data, the BS performs SIC in all slots in which that user has transmitted a packet [2].[1] This process repeats and the decoding at the BS proceeds in iterations.

Algorithm 1: Performance Evaluation of C-IRSA

```
Input: $T, M, \rho_0, \mathbf{G}, k_{\max}, \mathcal{A} = \{i \in [M] | |h_i|^2 \geq \nu\}$
Initialize: $\mathcal{S}_1 = [M]$
for $k = 1, 2, \ldots, k_{\max}$ do
    for $t = 1, 2, \ldots, T$ do
        Evaluate the SINR $\rho_{ti}^k,\ \forall i \in \mathcal{S}_k$ from (3)
        If $\rho_{ti}^k \geq \gamma_{\text{th}}$, remove user $i$ from $\mathcal{S}_k$ and
          perform SIC in all slots where $g_{ti} = 1$
    end
end
Output: $\mathsf{PLR} = |\mathcal{S}_{k_{\max}}|/M,\ \mathcal{T} = M(1 - \mathsf{PLR})/T,$
    $\mathsf{PLR}_a = |\mathcal{A} \cap \mathcal{S}_{k_{\max}}|/|\mathcal{A}|.$
```

We use the SINR threshold model to abstract the decodability of any packet: a packet can be decoded correctly if and only if its SINR is above a threshold $\gamma_{\text{th}} \geq 1$ [3], [7]. To evaluate the performance of C-IRSA with the SINR threshold model, we first compute the SINRs achieved by all the users in all the slots in any decoding iteration. If there is a user with SINR $\geq \gamma_{\text{th}}$ in some slot, we consider that packet as successfully decoded and remove the contribution of that user's packet from all other slots in which that user has transmitted a replica [6]. We then proceed to the next decoding iteration and recompute the SINRs for all users yet to be decoded. This process stops when no additional users are decoded in two successive iterations. The throughput $\mathcal{T}$ is calculated as the number of correctly decoded unique packets divided by the number of slots.

The calculation of the SINR of the users is as follows. We define $\mathcal{S}_k$ as the set of users not decoded upto the kth iteration with $\mathcal{S}_k^m \triangleq \mathcal{S}_k \backslash \{m\}$ and $\mathcal{S}_1 = [M]$. We can write the received signal in the tth slot in the kth decoding iteration as

$$y_t^k = \textstyle\sum_{i \in \mathcal{S}_k} \sqrt{\rho_0} a_i g_{ti} h_i x_i + n_t. \tag{2}$$

In order to decode the mth user, we first compute the processed signal $\tilde{y}_{tm}^k \triangleq h_m^* y_t^k$, which can be written as

$$\tilde{y}_{tm}^k = \sqrt{\rho_0} a_m g_{tm} |h_m|^2 x_m + \textstyle\sum_{i \in \mathcal{S}_k^m} \sqrt{\rho_0} a_i g_{ti} h_m^* h_i x_i + h_m^* n_t,$$

where the first term $T_1 \triangleq \sqrt{\rho_0} a_m g_{tm} |h_m|^2 x_m$ is the desired signal, the second term $T_2 \triangleq \sum_{i \in \mathcal{S}_k^m} \sqrt{\rho_0} a_i g_{ti} h_m^* h_i x_i$ is the MUI, and $T_3 \triangleq h_m^* n_t$ is noise. Since noise is uncorrelated with the other terms and the data streams of distinct users are uncorrelated, the terms T_1, T_2, and T_3 are all uncorrelated with each other. The power in the received signal is a sum of the powers of the terms. Thus, the SINR, ρ_{tm}^k, of the mth user

[1] The set of slots where the user's packet is repeated can be included in the header of the packet.

in the tth slot in the kth iteration, can be computed as

$$\rho_{tm}^{k}=\frac{\rho_0 a_m g_{tm}|h_m|^2}{1+\sum_{i\in\mathcal{S}_k^m}\rho_0 a_i g_{ti}|h_i|^2}. \tag{3}$$

The performance of C-IRSA can now be computed as detailed in Alg. 1. Here, the decoding proceeds for $k_{\max}$ iterations, and the output is the system throughput, $\mathcal{T}$, the packet loss rate (PLR) of the active users, PLR_a, and the system PLR, PLR.

Remarks: The threshold ν can be declared by the BS during pilot transmission based on the system load; the optimal choice of ν is discussed in the sequel. Also, we ignore channel estimation errors in determining whether the user remains active/inactive and in calculating the SINR in (3). However, it is straightforward to include these effects using the results in [3]. Finally, the BS can determine which users are active in each frame, for example, using the user activity detection (UAD) algorithm presented in [4]. It is shown in [4] that a short pilot transmission from the users for channel estimation at the BS is also sufficient for accurate UAD.

III. Theoretical Analysis of C-IRSA

In the previous section, we described an *empirical* approach to evaluate the performance of C-IRSA, given by Alg. 1. We now characterize the theoretical performance of C-IRSA using density evolution (DE) [2], [3]. SIC-based decoding can be viewed as message passing on a bipartite graph [6], and thus C-IRSA can be decoded on graphs. The bipartite graph is made up of the user nodes on one side, the slot nodes on the other side, and the edges between them. An edge connects a user node to a slot node if and only if that user has transmitted a packet in that corresponding slot. DE is applicable as M_a and $T\to\infty$ with a fixed $L_a=M_a/T$ [7]. Hence, we describe the DE process in terms of only the active load L_a. Due to lack of space, we only outline the high-level steps in the analysis here. Detailed discussion of the DE technique can be found in several references [2], [3], [6], [7].

The repetition factor of a user is the number of replicas the user has transmitted in a given frame, whereas the collision factor of a slot is the number of packets that have collided in that slot. The *node-perspective user degree distribution* is the set of probabilities $\{\phi_d\}_{d=2}^{d_{\max}}$, where ϕ_d is the probability that a user has a repetition factor d; with minimum and maximum repetition factors of 2 and $d_{\max}$, respectively. The *edge-perspective user degree distribution* is the set of probabilities $\{\lambda_d\}_{d=2}^{d_{\max}}$, where $\lambda_d=d\phi_d/\phi'(1)$ is the probability that an edge is connected to a user with repetition factor d. The corresponding polynomial representations of the node- and edge- perspective user degree distributions are

$$\phi(x)=\textstyle\sum_{d=2}^{d_{\max}}\phi_d x^d,\quad \lambda(x)=\sum_{d=2}^{d_{\max}}\lambda_d x^{d-1}, \tag{4}$$

respectively. The average repetition factor is $\bar{d}\triangleq\sum_d d\phi_d$.

The degree distributions defined above are now used to find a pair of interdependent *failure probabilities* denoted by "p_i" and "q_i" in the ith decoding iteration. The user and slot nodes exchange failure messages along an edge when a decoding failure happens, i.e., when that user has not been decoded in that slot in the current decoding iteration. The probability that an edge carries a failure message from a slot node to a user node is denoted by p_i, whereas the probability that an edge carries a failure message from a user node to a slot node is denoted by q_i. Using the edge-perspective user degree distribution, the failure probability q_i is calculated as

$$q_i=\textstyle\sum_{d=2}^{d_{\max}}\lambda_d p_{i-1}^{d-1}=\lambda(p_{i-1}). \tag{5}$$

Here, the probability that an edge carries a failure message in the ith iteration given that it is connected to a user node with repetition factor d is p_{i-1}^{d-1}. If all the other $d-1$ incoming edges to that user node carry failure messages in the previous iteration, then the edge will carry a failure message from that user node in the ith iteration. The failure probability p_i is calculated as in [3], [7] as

$$p_i=1-e^{-L_a\bar{d}q_i}\textstyle\sum_{r=1}^{\infty}\theta_r(L_a\bar{d}q_i)^{r-1}/(r-1)!\triangleq f(q_i). \tag{6}$$

Here, θ_r is the probability that a reference packet gets decoded in any iteration in a slot of degree r using only intra-slot SIC [7]. Intra-slot SIC refers to interference cancellation within the same slot a user is decoded in, whereas inter-slot SIC refers to interference cancellation in a different slot. We now describe the evaluation of θ_r, which is the crucial step in computing the throughput.

Theorem 1. *For the Rayleigh block-fading channel with an SNR of ρ_0, a censor threshold ν, and a decoding threshold γ_{th}, the probability that a reference packet gets decoded in a slot of degree r using only intra-slot SIC, can be obtained as*

$$\theta_r=\sum_{k=1}^{r}\frac{\exp(r\nu-(r-k)\nu\bar{\gamma}_{\text{th},k}-\rho_0^{-1}(\bar{\gamma}_{\text{th},k}-1))}{r\ \bar{\gamma}_{\text{th},k}^{\,r-(k+1)/2}}, \tag{7}$$

where $\bar{\gamma}_{\text{th},k}=(1+\gamma_{\text{th}})^k$, and $\nu\le\rho_0^{-1}\gamma_{\text{th}}$.

Proof. See Appendix A. ■

Remark: When $\nu=0$, i.e., there is no censoring, the expression for θ_r matches with the results by Clazzer et al. [7].

In DE, $q_i=\lambda(p_{i-1})$ and $p_i=f(q_i)$ are calculated recursively as functions of each other using (5) and (6), with either $q_0=1$ or $p_0=f(1)$. At the end of decoding, the failure probability is $p_\infty=\lim_{i\to\infty}p_i$. The probability that a packet from a user with repetition factor d does not get decoded at all is $(p_\infty)^d$. Therefore, the asymptotic packet loss rate of the active users (PLR_a), which is the fraction of packets of active users that are not decoded at the BS, is calculated as

$$\mathsf{PLR}_a=\phi(p_\infty)=\textstyle\sum_{d=2}^{d_{\max}}\phi_d(p_\infty)^d. \tag{8}$$

We denote the cumulative distribution function (CDF) and the complementary CDF of the exponential distribution (of $|h|^2\sim\exp(1)$) evaluated at x by $\mathrm{F}(x)$ and $\bar{\mathrm{F}}(x)\triangleq 1-\mathrm{F}(x)$, respectively. The active load L_a of the system is $L_a=L\bar{\mathrm{F}}(\nu)$. Since the fraction of censored users is $\mathrm{F}(\nu)$, the effective PLR of the system (including censored users) can be calculated as

$$\mathsf{PLR}=\mathrm{F}(\nu)+\bar{\mathrm{F}}(\nu)\mathsf{PLR}_a. \tag{9}$$

The throughput $\mathcal{T}$ of the users in the system can now be obtained from the asymptotic PLR as

$$\mathcal{T} = L(1 - \mathsf{PLR}) = L_a(1 - \mathsf{PLR}_a). \tag{10}$$

The iterations $p_i = f(\lambda(p_{i-1}))$ converge to $p_\infty = 0$ if the active load $L_a < L_a^*$, asymptotically [2], [3]. Here, L_a^* is called the *active inflection load* of the system, and it corresponds to a *system inflection load* of $L^* = L_a^*/\bar{\mathrm{F}}(\nu)$, with a threshold ν. For $L_a < L_a^*$, since $p_\infty = 0$, we have $\mathsf{PLR}_a = 0$, $\mathsf{PLR} = \mathrm{F}(\nu)$, and $\mathcal{T} = L\bar{\mathrm{F}}(\nu) = L_a$. For any $L_a \geq L_a^*$, PLR_a does not converge to 0, and $\mathcal{T}$ decreases monotonically with L_a. Also, from (9), we see that $\mathsf{PLR} \geq \mathrm{F}(\nu)$, and thus, $\mathcal{T} \leq L\bar{\mathrm{F}}(\nu)$.

1) Choice of Threshold: In order to choose ν, we first choose a target PLR for the active users, $\mathsf{PLR}_{a,\text{tgt}}$, which is a maximum permissible PLR among the active users. Let L_{tgt} be the *target load*, which is the minimum L at which the system achieves an active PLR of $\mathsf{PLR}_{a,\text{tgt}}$, with $\nu = \rho_0^{-1}\gamma_{\text{th}}$. At L_{tgt}, the active load is $L_a = L_{\text{tgt}}\bar{\mathrm{F}}(\rho_0^{-1}\gamma_{\text{th}})$, with a corresponding throughput of $\mathcal{T}_{\text{tgt}}$. For a load $L \geq L_{\text{tgt}}$, we wish to continue to operate at the same PLR of $\mathsf{PLR}_{a,\text{tgt}}$, to keep the throughput fixed at $\mathcal{T}_{\text{tgt}}$. This can be done by ensuring the same active load L_a at L and L_{tgt}. Thus, we need to choose ν such that

$$L_a = L\bar{\mathrm{F}}(\nu) = L_{\text{tgt}}\bar{\mathrm{F}}(\rho_0^{-1}\gamma_{\text{th}}). \tag{11}$$

Since $\bar{\mathrm{F}}(x) = \exp(-x)$, we obtain

$$\nu = \log(L/L_a) = \log(L/L_{\text{tgt}}) + \rho_0^{-1}\gamma_{\text{th}}. \tag{12}$$

The above is valid when $L \geq L_a$ or $L \geq L_{\text{tgt}}$. When $L < L_{\text{tgt}}$, as we will see in Fig. 3, the threshold that maximizes the throughput occurs at $\nu = \rho_0^{-1}\gamma_{\text{th}}$. An intuitive reason for this is that the probability of decoding a user, if that user was the only one transmitting in a slot, is $\theta_1 = \Pr(|h_1|^2 \geq \rho_0^{-1}\gamma_{\text{th}} \mid |h_1|^2 \geq \nu)$ $= \exp(\nu - \rho_0^{-1}\gamma_{\text{th}}) \cdot \mathbb{1}\{\nu \leq \rho_0^{-1}\gamma_{\text{th}}\} + \mathbb{1}\{\nu > \rho_0^{-1}\gamma_{\text{th}}\}$, when the threshold is ν. So if we set $\nu > \rho_0^{-1}\gamma_{\text{th}}$ or $\nu < \rho_0^{-1}\gamma_{\text{th}}$, we are censoring more or fewer users than required, respectively. Thus, the optimal choice of the censor threshold is given by the function $g(\cdot,\cdot)$ defined as

$$\nu = g(L, L_{\text{tgt}}) \triangleq \begin{cases} \rho_0^{-1}\gamma_{\text{th}}, & L < L_{\text{tgt}}, \\ \log(L/L_{\text{tgt}}) + \rho_0^{-1}\gamma_{\text{th}}, & L \geq L_{\text{tgt}}. \end{cases} \tag{13}$$

For $\nu = \rho_0^{-1}\gamma_{\text{th}}$, the system inflection load is $L^* = L_a^*/\bar{\mathrm{F}}(\rho_0^{-1}\gamma_{\text{th}})$. For $L_{\text{tgt}} < L^*$, the set of functions $\{g(\cdot,\cdot)\}$ achieve $\mathsf{PLR}_a \leq \mathsf{PLR}_{a,\text{tgt}}$ among the set of active users. In practice, we set a low target PLR of $\mathsf{PLR}_{a,\text{tgt}} \approx 10^{-3}$ or 10^{-4}.

IV. Numerical Results

In this section, we illustrate the performance of C-IRSA via Monte Carlo simulations. We also show how C-IRSA helps to overcome packet losses both due to poor CSI as well as due to MUI, as the load increases. In every simulation, we generate independent realizations of the channels and the access pattern matrix, and empirically evaluate the throughput of C-IRSA using Alg. 1. We also evaluate the theoretical throughput of C-IRSA as discussed in Sec. III and provide insights into the impact of various system parameters on the performance. The results are presented for 10^4 Monte Carlo runs, SNR $\rho_0 = 10$ dB, SINR threshold $\gamma_{\text{th}} = 10$ [3]. We use the truncated Soliton distribution [12] $\phi(x) = 0.625x^2 + 0.25x^3 + 0.125x^4$ to generate the repetition factors of the users [3].[2] The repetition factor d_i is used to form the access vector for the ith user, by uniformly randomly choosing d_i slots from T slots without replacement [2]. The packet replicas are transmitted in these d_i slots. For the empirical results, the number of users M is computed based on L as $M = \lfloor LT \rceil$; whereas the theoretical performance is dependent only on L, as described in Sec. III.

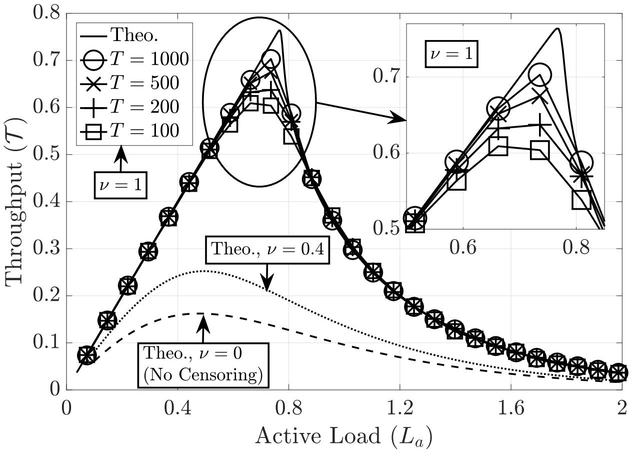

Fig. 1: Impact of T on the throughput.

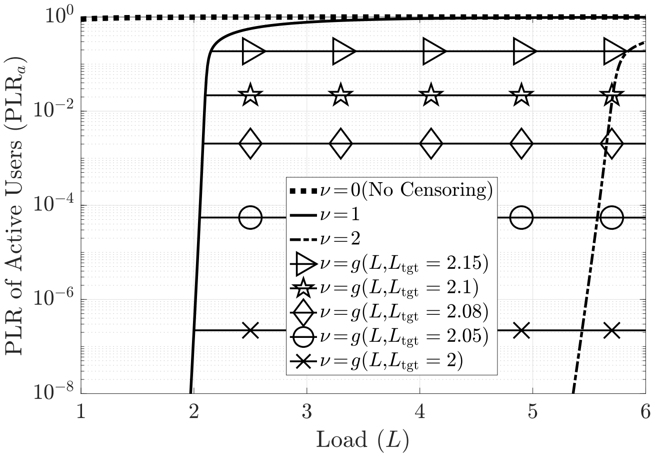

Fig. 2: Choice of target load L_{tgt} using theoretical PLR_a.

Fig. 1 shows the impact of T on the empirical throughput with $\nu = \rho_0^{-1}\gamma_{\text{th}} = 1$. The theoretical asymptotic throughput curves for $\nu = 0, 0.4$, and 1, obtained via DE, are also shown. The curves linearly increase till a peak, after which they drop quickly to zero as the system becomes MUI-limited. The asymptotic $\mathcal{T}$ is maximized at $L_a^* = \mathcal{T} = 0.76$, for $\nu = 1$. The linear increase in $\mathcal{T}$ marks the region in which $\mathsf{PLR}_a = 0$ [2]: when $L_a \leq 0.76$, all active users are decoded. Conventional IRSA corresponds to no censoring ($\nu = 0$). At $L_a = 0.4$, $\nu = 0$ achieves $\mathcal{T} = 0.15$, whereas $\nu = 1$ achieves full throughput

[2] We do not optimize the repetition distribution in this work since the goal is to evaluate the impact of censoring.

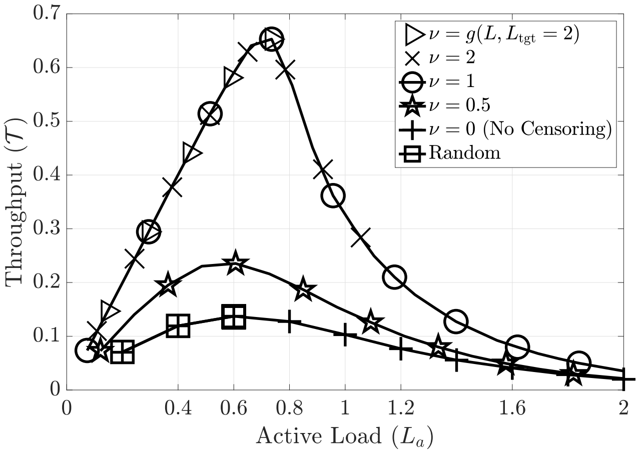

Fig. 3: Effect of active load L_a on $\mathcal{T}$.

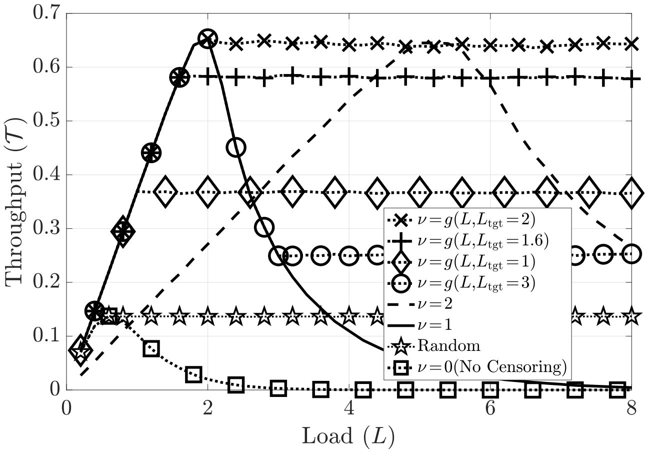

Fig. 4: Impact of threshold ν on $\mathcal{T}$.

of $\mathcal{T} = L_a = 0.4$. The asymptotic throughput dramatically improves as ν is increased from 0 to 1, because users with poor channel states are self-censored. Even with a little amount of censoring, C-IRSA performs better than IRSA. Thus, C-IRSA helps overcome packet losses due to both poor CSI and MUI.

We have seen that the choice of the threshold ν must be such that $L_a \leq L_a^*$. In Fig. 2, we depict the influence of the choice of the target load, $L_{\rm tgt}$, using the asymptotic active PLR, PLR_a. The PLR is close to 1 with no censoring. The PLR_a of the system increases with L for $\nu = 0, 1$, and 2, and becomes 1 at high loads. The curves with $\nu = g(L, L_{\rm tgt})$ follow the performance of $\nu = \rho_0^{-1}\gamma_{\rm th} = 1$ upto a load of $L = L_{\rm tgt}$, and beyond that PLR_a stays constant at every load. Fixing a $\mathsf{PLR}_{a,\rm tgt}$ yields the choice of $L_{\rm tgt}$ and the corresponding threshold $\nu = g(L, L_{\rm tgt})$. The asymptotic PLR increases very quickly around the inflection load L^*. In practice, however, choosing $L_{\rm tgt} = 0.9L^*$ or $0.8L^*$ works well.

In Fig. 3, we show the effect of the active load L_a on the empirical throughput $\mathcal{T}$, with $T = 250$. Conventional IRSA (no censoring, i.e., $\nu = 0$) achieves very low throughputs since the system is highly interference limited. Similar to the previous plot, where the theoretical throughput increased with increase in ν, the empirical throughput also increases with an increase in from $\nu = 0$ to $\nu = \gamma_{\rm th}/\rho_0 = 1$. For $\nu \geq \gamma_{\rm th}/\rho_0$, the throughput of the system stays constant with respect to the active load and the system achieves the same throughput for $\nu = 2$ as for $\nu = 1$. From the plot, we also see that we should choose a threshold ν such that we always operate the system at active load of $L_a \leq L_a^* = 0.65$. Also, by optimally choosing the threshold using $\nu = g(L, L_{\rm tgt})$ as described in Sec. III-1, we can obtain the same throughput as that obtained with $\nu = 1$. Note that in the MUI-limited regime, the PLR of IRSA is nonzero, and both users with poor channel states as well as users who collide with many users cannot be decoded correctly. Censoring improves the performance of the system on both counts by choosing users whose packets are more likely to be decoded correctly as well as reducing the number of collisions.

So far, we have observed that both the theoretical and empirical throughputs are maximized at $\nu = \rho_0^{-1}\gamma_{\rm th}$ for every L_a. We now study the effect of censoring and the system load L on the empirical throughput in Fig. 4, with $T = 250$. With $\nu = 0$, i.e., no censoring, the throughput of IRSA becomes zero at $L = 3$. With $\nu = \rho_0^{-1}\gamma_{\rm th} = 1$, the throughput of the system increases linearly with load upto $\mathcal{T} = 0.65$ at $L = 2$, and beyond that, the throughput drops to zero. This is also observed with $\nu = 2$, which achieves a peak throughput of $\mathcal{T} = 0.65$ at $L = 5$. The linearity of the curve upto $L = 5$ indicates that too many users are self-censoring, and we could reduce ν. For $\nu = 1$, we have $\mathsf{PLR}_a = 0$ and $\mathsf{PLR} = \bar{\mathrm{F}}(1)$ upto $L = 2$; for $\nu = 2$, we have $\mathsf{PLR}_a = 0$ and $\mathsf{PLR} = \bar{\mathrm{F}}(2)$ upto $L = 5$. Thus, we could choose ν for every L such that we obtain an envelope of all curves for $\nu \geq 1$, which yields the same performance as that of the curve marked $\nu = g(L, L_{\rm tgt} = 2)$.[3] All the curves marked $\nu = g(L, L_{\rm tgt})$ follow the performance of $\nu = 1$ upto $L_{\rm tgt}$, beyond which $\mathcal{T}$ stays constant for every L. Since $L^* = 2$, choosing $L_{\rm tgt} = 3$ is not preferred since the system is operating at a high PLR. Choosing $L_{\rm tgt} = 1, 1.6$, and 2 all yield $\mathsf{PLR}_a = 0$ at all L since the active load $L_a \leq 0.65$. We thus choose $L_{\rm tgt} = 2$ to maximize $\mathcal{T}$, which can be obtained from our analysis as $L_{\rm tgt} = L_a^*/\bar{\mathrm{F}}(\rho_0^{-1}\gamma_{\rm th}) = 0.65/\bar{\mathrm{F}}(1) = 2$. Since the DE curves are achieved asymptotically, in practice, we back off from $L_{\rm tgt}$ by 10% to 20% to $L_{\rm tgt} = 1.8$ or 1.6, to achieve zero PLR_a at all L. At high L, we see that C-IRSA with $L_{\rm tgt} \leq L^*$ operates with $\mathcal{T} = 0.65$, whereas conventional IRSA has $\mathcal{T} = 0$. Thus, the system can be operated at its maximum potential in C-IRSA compared to vanilla IRSA which has zero throughput.

1) Impact of random censoring: The censoring of users can be done in a random fashion as opposed to CSI-based censoring: users independently participate in each frame with probability p_a, and self-censor with probability $1 - p_a$. This yields an active load of $L_a = Lp_a$. The optimal random censoring can be done by choosing $p_a = L_a^*/L$, since this ensures that $L_a = L_a^*$. The curve marked "Random" in Fig. 3 uses random censoring and achieves the same throughput as

[3] The theoretical throughputs for Figs. 3 and 4 match the above observations. Due to lack of space, we have not included them. Also, the results are presented for $\rho_0^{-1}\gamma_{\rm th} = 1$. The trends are similar for any other $\rho_0^{-1}\gamma_{\rm th}$, and $\mathcal{T}$ is maximized at $\nu = \rho_0^{-1}\gamma_{\rm th}$ for every L_a.

conventional IRSA for every $p_a \in (0, 1]$. For the same active load L_a, the channel states of the uncensored users with CSI-based censoring are better than the channel states of the active users with random censoring. With optimal random censoring, in order to operate the system at $\mathcal{T}_{\text{tgt}} = 0.15$ at $L_a^* = 0.6$, we need to choose $p_a = \min\{1, 0.6/L\}$. With this choice of p_a, we obtain the curve marked "Random" in Fig. 4, which achieves $\mathcal{T} = 0.15$ at all $L \geq 0.6$. Thus, the optimal CSI-based censoring in C-IRSA achieves a peak throughput of $\mathcal{T} = 0.65$ whereas optimal random censoring in IRSA has a peak throughput of $\mathcal{T} = 0.15$, an over 4× improvement.

V. Conclusion

In this work, we proposed a variant of IRSA, called C-IRSA, to overcome the performance degradation of IRSA at high loads. In C-IRSA, users self-censor based on their CSI, and the protocol retains the fully distributed, random access nature of IRSA. We derived closed-form expressions for the success probability θ_r for CSI-based censoring, and theoretically characterized the asymptotic performance of C-IRSA. Our analysis allows us to determine the optimal choice of the censor threshold ν, with which the PLR of the active users can be driven close to zero and yields the highest possible throughput. The results showed that we can achieve a 4× improvement in C-IRSA compared to optimal random censoring. Future work could account for CSI and load estimation errors, optimize the repetition distribution under C-IRSA, and also include a proportional fairness mechanism for users with poor CSI.

Appendix A: Proof of Theorem 1

We now characterize θ_r, which is the probability of decoding a reference packet in a *single slot* where r users have transmitted their packets. Since there is only one slot under consideration, users are decoded via intra-slot SIC. The reference packet is one of the r packets, and it gets decoded only if the packets having a higher SINR get successfully decoded first. Hence, they must also satisfy the SINR $\geq \gamma_{\text{th}}$ constraint. Thus, θ_r is the probability that the reference packet and the packets with higher SINRs all get decoded.

We denote the set of active users who have not yet been decoded in the first $k-1$ intra-slot decoding iterations by $\mathcal{S}_k$, and $\mathcal{S}_k^m \triangleq \mathcal{S}_k \setminus \{m\}$, with $\mathcal{S}_1 = [r]$. The SINR of the mth user in the kth intra-slot decoding iteration, ρ_m^k, is calculated as $\rho_m^k = |h_m|^2/(\rho_0^{-1} + \sum_{i \in \mathcal{S}_k^m}^r |h_i|^2)$. Let $\rho_{\max}^k$ denote the SINR of the user with the highest SINR in the kth intra-slot decoding iteration, calculated as $\rho_{\max}^k = \max_{m \in \mathcal{S}_k} \ \rho_m^k$. Let s be the index of the intra-slot decoding iteration in which the reference packet is decoded, with $1 \leq s \leq r$. Thus, θ_r is calculated as $\theta_r = \Pr(\rho_{\max}^1 \geq \gamma_{\text{th}}, \rho_{\max}^2 \geq \gamma_{\text{th}}, \ldots, \rho_{\max}^s \geq \gamma_{\text{th}})$. Since the reference packet is tagged uniformly at random from the users, the reference packet is equally likely to get decoded in any decoding iteration. We denote the probability that the k packets with the highest SINRs across decoding iterations all exceed the threshold γ_{th} by $\theta_{rk} \triangleq \Pr(\rho_{\max}^1 \geq \gamma_{\text{th}}, \rho_{\max}^2 \geq \gamma_{\text{th}}, \ldots, \rho_{\max}^k \geq \gamma_{\text{th}})$. We can calculate θ_r using θ_{rk} as $\theta_r = (\sum_{k=1}^r \theta_{rk})/r$. Without loss of generality, let the channels of the users be ordered as $|h_1|^2 \geq |h_2|^2 \geq \ldots \geq |h_r|^2$. Now,

$$\theta_{rk} = \Pr\left(\frac{|h_1|^2}{\rho_0^{-1} + \sum_{i=2}^r |h_i|^2} \geq \gamma_{\text{th}}, \frac{|h_2|^2}{\rho_0^{-1} + \sum_{i=3}^r |h_i|^2} \geq \gamma_{\text{th}}, \right.$$
$$\left. \ldots, \frac{|h_k|^2}{\rho_0^{-1} + \sum_{i=k+1}^r |h_i|^2} \geq \gamma_{\text{th}} \middle| |h_j|^2 \geq \nu, \forall j \in [r]\right). \quad (14)$$

The above is a conditional probability, conditioned on $|h_j|^2 \geq \nu$, since we are considering only uncensored users. Thus, θ_{rk} from (14) can be calculated equivalently as

$$\theta_{rk} = \Pr(t_1 \geq \gamma_{\text{th}}(\rho_0^{-1} + \textstyle\sum_{i=2}^r t_i), t_2 \geq \gamma_{\text{th}}(\rho_0^{-1} + \sum_{i=3}^r t_i),$$
$$\ldots, t_k \geq \gamma_{\text{th}}(\rho_0^{-1} + \textstyle\sum_{i=k+1}^r t_i)).$$

Here, t_i is a random variable follows a *truncated exponential* distribution with the density function $\mathrm{f}(t) = \exp(\nu - t) \cdot \mathbb{1}\{\nu \leq t < \infty\}$. Assuming $\nu \leq \rho_0^{-1}\gamma_{\text{th}}$, with $\bar{\gamma}_{\text{th},i} = (1 + \gamma_{\text{th}})^i$, θ_{rk} can be calculated as

$$\theta_{rk} = e^{r\nu} \int_\nu^\infty e^{-t_r} \mathrm{d}t_r \int_\nu^\infty e^{-t_{r-1}} \mathrm{d}t_{r-1} \cdots \int_\nu^\infty e^{-t_{k+1}} \mathrm{d}t_{k+1}$$
$$\times \int_{\gamma_{\text{th}}(\rho_0^{-1} + \sum_{i=k+1}^r t_i)}^\infty e^{-t_k} \mathrm{d}t_k \cdots \int_{\gamma_{\text{th}}(\rho_0^{-1} + \sum_{i=2}^r t_i)}^\infty e^{-t_1} \mathrm{d}t_1$$
$$= \frac{\exp(r\nu - (r-k)\nu\bar{\gamma}_{\text{th},k} - \rho_0^{-1}\gamma_{\text{th}}(\sum_{i=1}^k \bar{\gamma}_{\text{th},i-1}))}{\bar{\gamma}_{\text{th},k}^{r-(k+1)/2}}. \quad (15)$$

Thus, we get

$$\theta_r = \sum_{k=1}^r \frac{\exp(r\nu - (r-k)\nu\bar{\gamma}_{\text{th},k} - \rho_0^{-1}(\bar{\gamma}_{\text{th},k} - 1))}{r \ \bar{\gamma}_{\text{th},k}^{r-(k+1)/2}}. \quad (16)$$

References

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, 2021.

[2] 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.

[3] C. R. Srivatsa and C. R. Murthy, "On the impact of channel estimation on the design and analysis of IRSA based systems," *IEEE Trans. Signal Process.*, vol. 70, pp. 4186–4200, 2022.

[4] ——, "User activity detection for irregular repetition slotted aloha based MMTC," *IEEE Trans. Signal Process.*, vol. 70, pp. 3616–3631, 2022.

[5] 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.

[6] E. E. Khaleghi, C. Adjih, A. Alloum, and P. Muhlethaler, "Near-far effect on coded slotted ALOHA," in *Proc. PIMRC*, Oct 2017.

[7] 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.

[8] C. R. Srivatsa and C. R. Murthy, "Throughput analysis of PDMA/IRSA under practical channel estimation," in *Proc. SPAWC*, July 2019.

[9] ——, "Performance analysis of irregular repetition slotted aloha with multi-cell interference," in *Proc. SPAWC*, July 2022.

[10] J. Choi and J. Ding, "Network coding for K-repetition in grant-free random access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, 2021.

[11] J. Ding and J. Choi, "SIC aided K-repetition for mission-critical MTC in cell-free massive MIMO," in *Proc. CSCN*, 2021.

[12] 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.