Online Backoff Control for NOMA-Enabled Random Access Procedure for Cellular Networks

Jun-Bae Seo , *Member, IEEE*, Bang Chul Jung , *Senior Member, IEEE*, and Hu Jin , *Senior Member, IEEE*

Abstract—Uplink power domain non-orthogonal multiple access (NOMA) random access (RA) system allows users to transmit their packet with one of target receive power (TRP) levels at the base station (BS). Then the BS can separate the users in simultaneous transmission with successive interference cancellation (SIC), when the users choose different TRP levels. It enjoys a higher throughput through supporting simultaneous transmissions that occur frequently in RA system. This letter examines how to incorporate uplink NOMA RA system with the existing Long-Term Evolution (LTE) RA procedure with minimal modifications. Furthermore, to optimize the proposed RA procedure, we propose an online control algorithm for backoff interval and analyze its performance particularly with focus on the average RA delay. The results show that the proposed RA procedure achieves $0.5896L$ of throughput for a total of L RA preambles and guarantees the average RA delay by $1/(0.5896L-\lambda)$ for Poisson traffic with mean λ.

Index Terms—No-northogonal multiple access, online control, random access procedure.

Fig. 1. RA procedure in LTE system.

I. INTRODUCTION

NON-ORTHOGONAL multiple access (NOMA) has been recently proposed as an uplink random access (RA) for the fifth (5G) generation cellular networks to improve the throughput [1]–[8]. Especially in [1]–[5], it requires user equipments (UEs) to control their transmit power such that their receive power at the base station (BS) can be one of the predefined target receive powers (TRP) in power-domain NOMA. The BS then decodes the received packets with successive interference cancellation (SIC) in the descending order of TRPs. Accordingly, by supporting simultaneous transmissions from UEs that may occur frequently, the throughput of
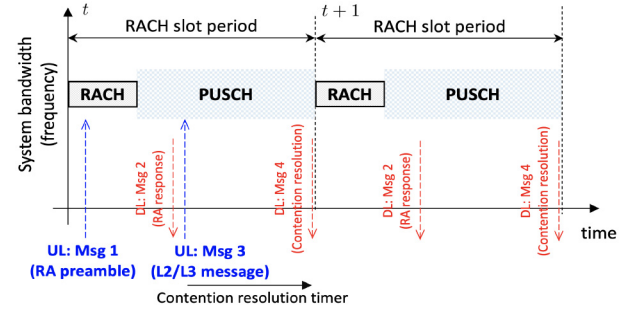
RA channel can be improved. So far there has been little consideration how uplink NOMA RA system can be related to the existing Long-Term Evolution (LTE) RA procedure. If a new system, e.g., uplink NOMA, would be implemented in a way of utilizing no part of the existing LTE system at all, it may increase the hardware cost of UEs that should access to both existing and new systems independently and capital expenditure (CAPEX) of network operators that employ both of them. It can be a cost-effective solution when uplink NOMA RA system can be seamlessly integrated into the existing LTE-RA procedure. This letter examines how to integrate uplink NOMA RA system into LTE RA procedure with a minor modification on the existing standard [9] especially in the medium access control (MAC) layer perspective. Our contributions can be summarized as follows:

- We propose uplink NOMA-LTE RA procedure, in which uplink NOMA RA system [1]–[3] can be seamlessly integrated into LTE RA procedure. The only modification on UE side is to use two TRP levels in transmitting the connection request message often denoted by Msg 3 in [9], whereas eNodeB, i.e., BS of LTE system, should be capable of SIC.
- To make the most out of NOMA-LTE RA procedure, we develop an online control algorithm for backoff interval that can adapt to network population and analyze its performance. In our analysis, the average RA delay is accurately estimated so that the average traffic load allowed to the system can be explicitly found to guarantee a target average RA delay.

It is notable that the uplink RA system of 5G New Radio (NR) networks employs the LTE RA procedure in Fig. 1 as well. The performance of our proposed procedure shall be compared with that of the existing LTE RA procedure (without NOMA) in Section IV.

II. SYSTEM MODEL

A. The Current RA Procedure

Time domain of the uplink of LTE system is divided into frames, each of which consists of 10 subframes numbered from 0 to 9, whereas one subframe is 1 msec long. The physical random access channel (PRACH) appears periodically in a frame, whereas the time period is determined by PRACH configuration index. Among a total of 64 configuration indices, when configuration index 3 is used, PRACH appears in subframe 1. As shown in Fig. 1, we define one RA slot period as a time period of PRACH [12], [13]. It starts with the beginning of PRACH and includes the following PUSCH.

When UEs have a packet to send, called backlogged, they first transmit a RA preamble to PRACH. The existing LTE-RA procedure is as follows:

- *Step 1:* Each backlogged UE chooses one of L RA preambles and transmits it to PRACH. This RA preamble is denoted by Msg 1 as shown in Fig. 1.
- *Step 2:* After receiving preambles from the backlogged UEs, eNodeB sends random access response (RAR) message to them over the downlink, which is denoted by Msg 2 in Fig. 1. Notice that the eNodeB allocates an uplink resource in PUSCH for each RA preamble transmitted. Therefore, if two or more UEs select the same RA preamble, they get the same uplink resource in PUSCH. The reason for this is that when multiple of identical RA preambles are received, the eNodeB can not help but treat them as multipath signals.
- *Step 3:* Upon RAR message reception, the UEs send Msg 3 to the allocated uplink resource in PUSCH and initiate a contention resolution timer during which they shall monitor PDCCH for feedback from the eNodeB. Since the UEs with the same RA preamble transmitted send this Msg 3 to the same uplink resource, their Msg 3 will be in collision at the eNodeB.
- *Step 4:* It can be envisioned that the eNodeB decodes successfully the connection request message from the UEs that have transmitted a unique RA preamble at Step 1. Then, the eNodeB sends the contention resolution message (Msg 4) to all successful UEs. On the other hand, those UEs who experience Msg 3 collision at the eNodeB cannot receive Msg 4. Once, the contention resolution timer expires, those collided UEs shall restart Step 1 after waiting a random time taken uniformly within the backoff interval specified by the eNodeB.

Before presenting our proposed RA procedure, we make the following assumptions. *Firstly,* we assume that the four-step RA procedure above is finished before the next PRACH starts as shown in Fig. 1. The contention resolution timer set in Step 3 also expires in synchronization with Msg 4 reception. Accordingly, when the UEs do not find themselves in Msg 4, due to the expiry of contention resolution timer they shall restart Step 1 after backoff. *Secondly,* when U_t denotes a backoff interval broadcast by the eNodeB for RA slot t, it is informed in RAR message at Step 2. LTE-RA procedure defines 13 values of backoff interval: 5, 10, 20, 30, 40, 60, 80, 120, 160, 240, 320, 480, 960, and 1920 (msec). These are also used in RA system of 5G NR. We assume that the backoff interval U_t is an integer multiple of RA slot periods in order to make the backoff interval fine-grained. *Thirdly,* at Step 3, the UEs may retransmit Msg 3 up to the maximum of H-ARQ retransmissions over several RA slots. We however assume that Msg 3 is transmitted only once. Note that if multiple UEs transmit Msg 3 at the same PUSCH resource, H-ARQ does not help due to collision. *Fourthly,* we assume that a UE can store only one packet which corresponds to one connection request message. When a UE has a packet to transmit, it will read U_t broadcast in the previous RA slot and take a backoff interval, which is known as delayed first transmission [13]. We *finally* consider a backoff algorithm that the eNodeB broadcasts a retransmission probability r_t every slot such that UEs (re)transmit their packet according to Bernoulli trial with probability r_t. To distinguish this from the proposed online control backoff algorithm with the backoff interval U_t, it is called probability-based backoff algorithm.

B. The Proposed RA Procedure: NOMA-LTE

In LTE-RA procedure, the UEs shall transmit Msg 3 at Step 3 by keeping the receive power of Msg 3 at the eNodeB a TRP so that the eNodeB can decode it [9]. The proposed RA procedure makes use of power-domain NOMA at Step 3 as follows: UEs transmit Msg 3 by randomly aiming at one of two TRP levels such as P_1 and P_2 for $P_1 > P_2$. If two UEs chance to transmit the same RA preamble at Step 1 and each of them transmits with P_1 and P_2, respectively, the eNodeB can decode their Msg 3 by using SIC with the following condition:

$$\frac{P_1}{P_2 + N_0} \geq \gamma \text{ and } \frac{P_2}{N_0} \geq \gamma, \tag{1}$$

where N_0 and γ denote additive white noise and the decoding threshold, respectively. When both UEs transmit Msg 3 with P_1 or P_2, the eNodeB can not decode both. Based on (1), we find the minimum of P_1 and P_2 respectively as

$$P_1 = (1 + \gamma)\gamma N_0 \quad \text{and} \quad P_2 = \gamma N_0. \tag{2}$$

The proposed RA procedure considers only two TRP levels, no matter how many UEs transmit the same RA preamble at Step 1. When more than two UEs transmit Msg 3, none of them can be successfully decoded according to (2). When a UE transmits a unique RA preamble at Step 1, since it does not know whether it is the only UE transmitting a specific RA preamble, it shall transmit Msg 3 with either P_1 or P_2. The reason that we assume only two TRP levels is that since a collision made by two UEs occurs most highly likely, resolving such a collision with two TRP levels is quite efficient in terms of throughput improvement per the power consumption. If more power levels are used, a collision made by more than two UEs could be resolved in expense of much larger power consumption. Since uplink NOMA enables eNodeB to decode successfully two Msg 3's, we can expect that the number of contention resolution at Step 4 increases.

III. ANALYSIS AND ONLINE CONTROL ALGORITHM

A. Performance Analysis

Let X_t be the number of backlogged UEs at the beginning of RA slot t, whereas A_t denotes the number of UEs newly joining the backlog between RA slot t and $t + 1$. The number of UEs making a successful RA is denoted by S_t; that is, the sum of UEs that (re)transmit a unique RA preamble, which leads to a single Msg 3 transmission in PUSCH, and UEs that make a successful transmission of Msg 3 in PUSCH using NOMA, at RA slot t. Over time, X_t develops as

$$X_{t+1} = X_t - S_t + A_t, \qquad (3)$$

where $S_t \in \{0, 1, \ldots, \min(2L, X_t)\}$. Let J_t be the number of backlogged UEs that (re)transmit their packet with one out of L RA preambles. In what follows, we first consider the probability-based backoff algorithm with $r_t = r$.

Proposition 1: Given $X_t = m$ backlogged UEs, the probability that a preamble is transmitted by j UEs can be approximated by a Poisson distribution with mean rm/L (UEs/slot):

$$\Pr[J_t = j | X_t = m] \approx \Lambda_j(m) = \frac{\left(\frac{rm}{L}\right)^j}{j!} e^{-\frac{rm}{L}}. \qquad (4)$$

Proof: The exact expression of $\Pr[J_t = j | X_t = m]$ is expressed as $\sum_{k=j}^{m} \binom{k}{j}\left(\frac{1}{L}\right)^j\left(1-\frac{1}{L}\right)^{k-j}\binom{m}{k}r^k(1-r)^{m-k}$. To approximate it, for $j = 0$, we can write

$$\sum_{k=0}^{m}\left(1-\frac{1}{L}\right)^k\binom{m}{k}r^k(1-r)^{m-k} = \left(1-\frac{r}{L}\right)^m \approx e^{-\frac{rm}{L}},$$

where we have used $\sum_{i=0}^{m}\binom{m}{i}a^i b^{m-i}$ and $(1-x)^m \approx e^{-x}$. For $j = 1$, we have

$$\sum_{k=1}^{m} k\left(\frac{1}{L}\right)\left(1-\frac{1}{L}\right)^{k-1}\binom{m}{k}r^k(1-r)^{m-k}$$

$$= \frac{r}{L}\sum_{k=1}^{m}\frac{m(m-1)!}{(k-1)!(m-1-(k-1)!}$$

$$\times \left(r\left(1-\frac{1}{L}\right)\right)^{k-1}(1-r)^{m-1-(k-1)}$$

$$= \frac{rm}{L}\left(1-\frac{r}{L}\right)^{m-1} \approx \frac{rm}{L}e^{-\frac{rm}{L}},$$

where we have assumed $e^{-\frac{r(m-1)}{L}} \approx e^{-\frac{rm}{L}}$ as m grows large. For $j = 2$, we also have

$$\sum_{k=2}^{m}\frac{k(k-1)}{2}\left(\frac{1}{L}\right)^2\left(1-\frac{1}{L}\right)^{k-2}\binom{m}{k}r^k(1-r)^{m-k}$$

$$= \frac{1}{2}\left(\frac{r}{L}\right)^2\sum_{j=2}^{m}\frac{m(m-1)(m-2)!}{(k-2)!(m-2-(k-2)!}$$

$$\times \left(r\left(1-\frac{1}{L}\right)\right)^{k-2}(1-r)^{m-2-(k-2)}$$

$$= \frac{m(m-1)}{2}\left(\frac{r}{L}\right)^2\left(1-\frac{r}{L}\right)^{m-2} \approx \frac{1}{2}\left(\frac{mr}{L}\right)^2 e^{-\frac{rm}{L}},$$

where we have assumed $m(m-1) \approx m^2$ as m grows large. Exercising this for $j = 3, 4, \ldots$, completes the proof. ∎

Before deriving an optimal U_t, let us assume that the eNodeB broadcasts a retransmission probability r_t instead of U_t. UEs then retransmit RA preambles based on Bernoulli trial with probability r_t in order to build an analytical model.

Proposition 2: For $X_t = m$, let r_m^* be the retransmission probability of maximizing S_t. It can be obtained as

$$r_m^* = \min\left(1, \sqrt{2}L/m\right). \qquad (5)$$

Proof: When transmitting Msg 3, the UEs transmit it with P_1 and P_2 based on probability p and $1 - p$, respectively. Let μ_m be the mean number of UEs that make a successful RA when the system has m backlogged UEs, i.e., $\mu_m = \mathbb{E}[S_t | X_t = m]$. It is also the mean rate since its unit is the number of UEs per RA slot. Using Proposition 1, we can find it as

$$\mu_m = L\left(\Lambda_1(m) + 2\binom{2}{1}p(1-p)\Lambda_2(m)\right)$$

$$= rme^{-\frac{rm}{L}}\left(1 + 2p(1-p)\frac{rm}{L}\right), \qquad (6)$$

where the term in the parenthesis in the first line implies the number of packets successfully (re)transmitted for one RA preamble. We mulitplies L due to the independence among RA preambles. Let us consider $\frac{d\mu_m}{dr}$:

$$\frac{d\mu_m}{dr} = me^{-\frac{rm}{L}}\left(1 + \frac{m}{L}(4p\overline{p} - 1)r - 2p\overline{p}\left(\frac{m}{L}\right)^2 r^2\right). \qquad (7)$$

Let r_m be the solution of $\frac{d\mu_m}{dr} = 0$. It can be expressed as

$$r_m = \frac{-\frac{m}{L}(4p\overline{p} - 1) \pm \sqrt{\left(\frac{m}{L}(4p\overline{p} - 1)\right)^2 + 8p\overline{p}\left(\frac{m}{L}\right)^2}}{-4p\overline{p}\left(\frac{m}{L}\right)^2}. \qquad (8)$$

Since $p = 0.5$ maximizes (6) with respect to p, after plugging it, we get the positive r_m as (5). Note that r_m^* in (5) is a unique maximizer of μ_m, if μ_m is a quasi-concave function of r. To show this, let us recall that a function $f(x)$ defined on an interval I is quasi-concave if there exists a number x^* such that $f(x)$ is nondecreasing on $\{x \in I : x \leq x^*\}$ and nonincreasing on $\{x \in I : x \geq x^*\}$. Thus, it is necessary to show $\frac{d\mu_m}{dr} > 0$ (viz., μ_m is increasing) for $0 \leq r \leq r_m^*$ and $\frac{d\mu_m}{dr} < 0$ (μ_m is decreasing) for $r \geq r_m^*$. First, notice that $\mu_m = 0$ for $r = 0$ and ∞. We get $\frac{d\mu_m}{dr} = me^{-\frac{rm}{L}}(1 + \frac{1}{\sqrt{2}}\frac{m}{L}r)(1 - \frac{1}{\sqrt{2}}\frac{m}{L}r)$ for $p = 0.5$, from which it can be seen that $\frac{d\mu_m}{dr} > 0$ for $0 \leq r \leq \frac{\sqrt{2}L}{m}$ and $\frac{d\mu_m}{dr} < 0$ for $r > \frac{\sqrt{2}L}{m}$. This completes the proof. ∎

We now convert the (geometric) probability-based backoff algorithm in Proposition 2 into backoff interval U_t as follows.

Proposition 3: Let U_m denote a backoff interval in terms of the number of RA slots when the system estimates $X_t = m$, which can be obtained as

$$U_m = \text{nint}\left(\sqrt{2}m/L\right), \qquad (9)$$

where $\text{nint}(x)$ is the nearest integer function that takes the integer closest to x.

Proof: Notice that retransmissions by (5) are geometrically distributed, while the proposed system utilizes a discrete uniform window U_m. For $X_t = m$, we match the mean interval of

RA attempts of two backoff algorithms (online control backoff algorithm with backoff interval U_t and probability-based backoff algorithm) as $\frac{1}{r_m^*} = \frac{U_m}{2} \Rightarrow r_m^* = \frac{2}{U_m}$. Plugging this into (5) and discretizing it with nint($\cdot$) yields (9). ∎

To measure the performance of the proposed system, let us consider the steady-state probability that the system has m backlogged UEs. Let us denote it by $\pi_m = \lim_{t \to \infty} \Pr[X_t = m]$ for $m \in \mathbb{Z}$. Suppose that $\mathbb{E}[A_t] = \lambda_m$ for $X_t = m$. We approximate the Markov process described by (3) as a generalize M/M/1 queueing process. Then, we can write a flow balance equation for π_m as

$$\lambda_m \pi_m = \mu_{m+1} \pi_{m+1} \Rightarrow \pi_{m+1} = \frac{\lambda_m}{\mu_{m+1}} \pi_m. \quad (10)$$

For $n \in \mathbb{Z}^+$, in terms of π_0, we can rearrange (10) as

$$\pi_n = \pi_0 \prod_{m=0}^{n-1} \frac{\lambda_m}{\mu_{m+1}} \text{ for } n \geq 1. \quad (11)$$

We get $\pi_0 = (1 + \sum_{n=1}^{\infty} \prod_{m=0}^{n-1} \frac{\lambda_m}{\mu_{m+1}})^{-1}$ by $\sum_{n=0}^{\infty} \pi_n = 1$.

The mean RA delay of the proposed RA system can be characterized as follows.

Proposition 4: For Poisson traffic with mean rate λ (UEs/RA slot), the average RA delay (RA slots) that r_m in (5) can guarantee is expressed as

$$\overline{d} = \frac{1}{0.5896L - \lambda} + 0.5, \quad (12)$$

where 0.5 is added to account for the RA slot synchronization delay.

Proof: Using (5), we obtain μ_m for $m \in \mathbb{Z}$ as

$$\mu_m = L(1 + \sqrt{2})e^{-\sqrt{2}} \approx 0.5896L. \quad (13)$$

When $\lambda_m = \lambda$, the system is reduced to an ordinary M/M/1 queueing system. Let us write the system utilization ρ as

$$\rho = \frac{\lambda_m}{\mu_{m+1}} = \frac{\lambda}{0.5896L}. \quad (14)$$

For a stable system we should have $\rho < 1$, which means $\lambda < 0.5896L$. Under this condition, from (11) we get $\pi_n = (1 - \rho)\rho^n$. Let $\overline{n}$ and $\overline{d}$ denote the average number of backlogged UEs and the mean access delay, respectively. Using (11), and (14), we can find $\overline{n} = \sum_{i=0}^{\infty} i\pi_i = \frac{\rho}{1-\rho} = \frac{\lambda}{0.5896L - \lambda}$. Using Little's result, i.e., $\lambda \overline{d} = \overline{n}$, we get (12). ∎

Two remarks can be made: First, RA system can be a queueing system with random order of service (ROS). Therefore, although the RA delay distribution (waiting time distribution) a queueing system with ROS is different from the system with first-come first serve (FCFS), its mean is same owing to Little's result. Second, although we characterize the average RA delay based on Proposition 2, the analytical results in (12) are compared with simulations in Section IV.

Proposition 5: When r_m^* in (5) is used, the system throughput τ for $\lambda < 0.5896L$ is λ. In other words, when $\lambda < 0.5896L$, all the arriving packets could be eventually served.

Proof: The system throughput is expressed as $\overline{\tau} = \sum_{m=1}^{\infty} \mu_m \pi_m$. When (5) is substituted into this expression and $\pi_m = (1 - \rho)\rho^m$ is used, we simply have $\overline{\tau} = \lambda$. Note that this is valid if $\rho < 1$, i.e., $\lambda < 0.5896L$. ∎

Algorithm 1 NOMA-LTE Online Control Algorithm

Initialize $\widetilde{X}_0 = 10$, $\lambda_0 = 1$, and $\alpha = 0.99$. Repeat the following steps at the beginning of each RA slot.
1: $\widetilde{\lambda}_t = \alpha \widetilde{\lambda}_{t-1} + (1 - \alpha)S_{t-1}$
2: $\widetilde{X}_t = \widetilde{X}_{t-1} + 0.4543L - 1.8685\mathbb{I} - S_{t-1} + \widetilde{\lambda}_t$
3: Broadcast $U_t = \text{nint}\left(\frac{\sqrt{2}\widetilde{X}_t}{L}\right)$

For comparison, let us consider the performance of the existing LTE-RA system, where NOMA is not used for Msg 3 transmission. Instead of (6), we can use

$$\mu_m = L\Lambda_1(m). \quad (15)$$

The maximizer of (15) with respect to r is $r_m = \frac{L}{m}$. As in Proposition 3, we have $U_m = \text{nint}(\frac{2m}{L})$. Furthermore, substituting r_m back into (15), we have $\mu_m = 0.3679L$ for $m \in \mathbb{Z}^+$. Let $\hat{\rho}$ and $\hat{\pi}_m$ be the system utilization of LTE-RA system without NOMA and the steady-state probability that the system has m backlogged UEs. We then have $\hat{\pi}_m = (1 - \hat{\rho})\hat{\rho}^m$ for $\hat{\rho} = \lambda/(0.3679L)$. The average RA delay of LTE RA system is obtained by $\overline{d} = \frac{1}{0.3679L - \lambda} + 0.5$.

B. Online Control Algorithm

The objective of the proposed algorithm is to online control the backoff interval such that Proposition 3 can be realized. The proposed algorithm is presented in Algorithm 1, which is a modification of the recursive pseudo Bayesian algorithm in [11] in order to take into account Msg 3 transmission with NOMA. In line 1, $\widetilde{\lambda}_t$ denotes an estimation on $\mathbb{E}[A_t]$ in (3). Let us go back to (3) and take the expectation on both sides. We then have $\mathbb{E}[X_{t+1}] = \mathbb{E}[X_t - S_t + A_t] = \mathbb{E}[X_t] - \mathbb{E}[S_t] + \mathbb{E}[A_t]$. As $t \to \infty$, for a stable system, i.e., $\lambda < 0.5896L$ in (14), we have $\mathbb{E}[X_{t+1}] = \mathbb{E}[X_t]$ (the process in steady-state). This yields $\lim_{t \to \infty} \mathbb{E}[A_t] = \lim_{t \to \infty} \mathbb{E}[S_t]$. Thus, we can make our estimation on $\mathbb{E}[A_t]$ as in line 1, using the first-order autoregressive model.

When $X_t = m$ in (3), the system shall broadcast $U_t = \text{nint}(\sqrt{2}X_t/L)$ in realizing Proposition 3. However, it is not possible for the system to know X_t exactly every t. We thus use $U_t = \text{nint}(\sqrt{2}\mathbb{E}[X_t]/L)$. In line 2, $\widetilde{X}_t$ denotes the estimation on $\mathbb{E}[X_t]$. In what follows, we derive the update equation in line 2.

To begin with, let us assume that the number of backlogged UEs obeys a Poisson distribution with mean β (UEs/RA slot):

$$P_n(\beta) = \frac{\beta^n}{n!}e^{-\beta}, \quad (16)$$

which plays the *a priori* distribution on the number of backlogged UEs as in [11], i.e., belief on the backlog size. As a comparison with Proposition 1, using the belief of (16), we can express the probability that j UEs (re)transmit with a specific RA preamble given $\mathbb{E}[X_t] = \beta$ as $\Pr[J_t = j|\mathbb{E}[X_t] = \beta] = \sum_{k=j}^{\infty} \sum_{i=j}^{k} \binom{i}{j} r^j (1-r)^{i-j} \binom{k}{i} (\frac{1}{L})^i (1 - \frac{1}{L})^{k-i} P_k(\beta) = \frac{(\frac{r\beta}{L})^j}{j!} e^{-\frac{r\beta}{L}}$. For the sake of brevity, we leave out the derivation of this.

Let $\mathbb{B}$ and $\mathbb{I}$ denote the number of backlogged UEs and the number of RA preambles not transmitted, called idle
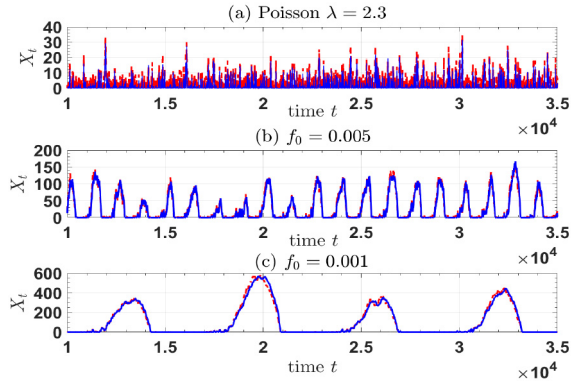
Fig. 2. Sample path of backlogged UEs X_t and the estimation $\widetilde{X}_t$.

Fig. 3. Average RA delay comparison of NOMA-LTE and LTE.

preambles, at RA slot t, respectively. By Bayesian rule, the conditional expectation of $\mathbb{B}$ given $\mathbb{I}$ can be expressed as

$$\mathbb{E}[\mathbb{B}|\mathbb{I}=k] = \sum_{n=0}^{\infty} n \Pr(\mathbb{B}=n|\mathbb{I}=k)$$
$$= \beta \left[(1-r) + r\left(1 - \frac{k}{L}\right)\left(1 - e^{-\frac{r\beta}{L}}\right)^{-1} \right], \quad (17)$$

where the second line is given as [11, eq. (19)]. Plugging $r = \sqrt{2}L/\beta$ into (17), we get

$$\mathbb{E}[\mathbb{B}|\mathbb{I}=k] = \beta + (0.4543L - 1.8685k). \quad (18)$$

When we subtract S_t, i.e., the number of UEs making a successful RA, from (18) and add λ_t to it, we have the update equation in line 2.

In LTE RA system without NOMA, plugging in $r = \frac{L}{\beta}$ and (17), we get $\mathbb{E}[\mathbb{B}|\mathbb{I}=k] = \beta + 0.582L - 1.582k$.

We replace the update equation in line 2 of Algorithm 1 with $\widetilde{X}_t = \widetilde{X}_{t-1} + 0.582L - 1.582\mathbb{I} - S_t + \widetilde{\lambda}_t$.

IV. NUMERICAL RESULTS

We build simulation with MATLAB and set each simulation run length to 10^6 RA slots and get the time-averaged result.

In Fig. 2 we observe how Algorithm 1 keeps track of X_t. In Fig. 2(a), L is set to 5 and Poisson process with mean rate $\lambda = 2.3$ (UEs/RA slot) is used as an input traffic, whereas Poisson process whose mean rate varies over time, $\lambda(t) = a\cos(f_0 t) + b$ is used for Figs. 2(b)–(c). We set $a = 0.2L$, $b = 0.45L$, $f_0 = 0.005$ in Fig. 2(b), and $f_0 = 0.001$ in Fig. 2(c). It can be seen that the proposed algorithm makes good estimation on X_t.

Fig. 3 presents the average RA delay over throughput (or traffic load). Symbols show simulation results and lines depict analytical results. Let us recall that analytical results are based on M/M/1 approximation with retransmission probability r_m^*, whereas in simulation, the system uses Algorithm 1 broadcasting U_t. It can be seen that M/M/1 approximation agrees well with simulations. The average RA delay of NOMA-LTE with $L = 5$ almost overlaps with that of LTE without NOMA. Note that NOMA-LTE yields $0.5896L = 2.948$ for $L = 5$ and LTE $0.3679L = 2.943$ for $L = 8$ so that the average delay $\overline{d}$ of the proposed system and LTE becomes almost identical. It can be found that applying NOMA in Msg 3 transmission, we can
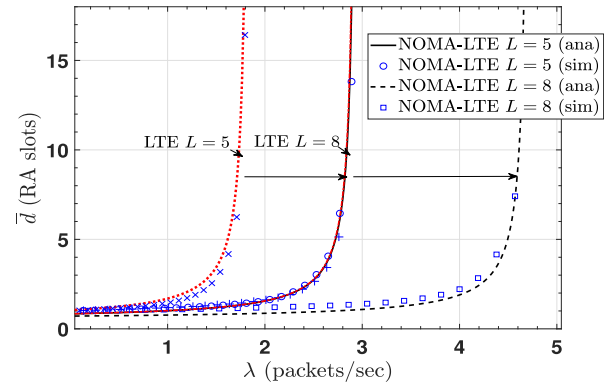
improve the maximum throughput $0.5896/0.3879 = 1.5955$ times. It can be concluded that if an objective of average RA delay $\overline{d}_o$ is given, the average traffic load λ should be limited by $\lambda < 0.5896L - 1/(\overline{d}_0 - 0.5)$.

V. CONCLUSION

This letter has shown how uplink NOMA RA system can be integrated into the existing LTE RA procedure: While keeping the current RA framework, NOMA RA can be seamlessly integrated by introducing one additional TRP level of Msg 3. In the proposed procedure, the eNodeB can separate Msg 3's using uplink NOMA. The timing estimation done with RA preambles might facilitate SIC for Msg 3. To optimize the proposed procedure, we have proposed an online control algorithm for backoff interval which improves the maximum throughput 1.595 times larger than the existing LTE RA procedure, whereas a target of the average RA delay can be guaranteed.

REFERENCES

[1] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.

[2] J.-B. Seo, B. C. Jung, and H. Jin, "Performance analysis of NOMA random access," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2242–2245, Nov. 2018.

[3] J.-B. Seo, B. C. Jung, and H. Jin, "Nonorthogonal random access for 5G mobile communication systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7867–7871, Aug. 2018.

[4] H. S. Jang, H. Lee, and T. Q. S. Quek, "Deep learning-based power control for non-orthogonal random access," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 2004–2007, Nov. 2019.

[5] D. Wang, Y. Qu, Y. Fu, Y. Yang, and Q. Chen, "A non-orthogonal random access scheme based on NB-IoT," *Wireless Pers. Commun.*, vol. 111, no. 99, pp. 2625–2639, 2020.

[6] Y. Yuan and C. Yan, "NOMA study in 3GPP for 5G," in *Proc. IEEE 10th Int. Symp. Turbo Codes Iterative Inf. Process. (ISTC)*, 2018, pp. 1–5.

[7] Y. Yuan, Z. Yuan, and L. Tian, "5G non-orthogonal multiple access study in 3GPP," *IEEE Commun. Mag.*, vol. 58, no. 7, pp. 90–96, Jul. 2020.

[8] "Study on non-orthogonal multiple access (NOMA) for NR, v16.0.0," 3GPP, Sophia Antipolis, France, Rep. TR 38.812, Dec. 2018.

[9] *Medium Access Control (MAC) Protocol Specification*, 3GPP Standard TS 36.321, 2020. [Online]. Available: www.3gpp.org

[10] Y. Xu, D. Cai, F. Fang, Z. Ding, C. Shen, and G. Zhu, "Outage analysis and power allocation for HARQ-CC enabled NOMA downlink transmission," in *Proc. GLOBECOM*, Dec. 2018, pp. 1–6.

[11] H. Jin, W. T. Toor, B. C. Jung, and J.-B. Seo, "Recursive pseudo-Bayesian access class barring for M2M communications in LTE systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8595–8599, Sep. 2017.

[12] S. Sesia, I. Toufik, and M. Baker, *LTE: The UMTS Long Term Evolution*. New York, NY, USA: Wiley, 2009.

[13] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A survey of traffic issues in machine-to-machine communications over LTE," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 865–884, Dec. 2016.