# Achieving Maximum Urgency-Dependent Throughput in Random Access

Yijin Zhang, *Senior Member, IEEE*, Aoyu Gong, Lei Deng, Yuan-Hsun Lo, *Member, IEEE*, Yan Lin, *Member, IEEE*, and Jun Li, *Senior Member, IEEE*

***Abstract*—Designing efficient random access is a vital problem for urgency-constrained packet delivery in uplink Internet of Things (IoT), which has not been investigated in depth so far. In this paper, we focus on unpredictable frame-synchronized traffic, which captures a number of scenarios in IoT communications, and generalize prior studies on this issue by considering a general ALOHA-like protocol, a general single-packet reception (SPR) channel, urgency-dependent throughput (UDT) based on a general urgency function, and the dynamic programming optimality. With a complete knowledge of the number of active users, we use the theory of Markov Decision Process (MDP) to explicitly obtain optimal policies for maximizing the UDT, and prove that a myopic policy is in general optimal. With an incomplete knowledge of the number of active users, we use the theory of Partially Observable MDP (POMDP) to seek optimal policies, and show that a myopic policy is in general not optimal by presenting a counterexample. Because of the prohibitive complexity to obtain optimal or near-optimal policies for this case, we propose two practical policies that utilize the inherent property of our MDP framework and channel model. Simulation results show that both outperform other alternatives. The robustness under relaxed system settings is also examined.**

***Index Terms*—Internet of Things, random access, stochastic optimal control, urgency constraint, delivery deadline**

## I. Introduction

### A. Background

RECENTLY there have been increasing demands for wireless technology to support real-time services in many application scenarios of the Internet of Things (IoT) [1]–[3], such as cooperative surveillance in sensor networks, cross-traffic assistance in vehicular-to-anything (V2X) networks, and process automation in industrial IoT. One common characteristic of packet delivery in these scenarios, termed as the *urgency constraint*, is that each packet is associated with a predefined strict deadline, contributes a fixed or higher reward when being earlier successfully delivered within its deadline, but is no longer useful after the deadline expiration.

Uplink IoT systems usually need to serve a large population of uncoordinated users with unpredictable traffic. For this canonical situation, achieving stringent performance targets under the urgency constraint is a challenging task due to the inherent coupling of delivery urgency and mutual interference. As a promising approach for this issue, ALOHA-like random access [4], [5] has been widely adopted in various IoT systems. Its basic idea is to allow the users to directly access certain radio resources in a contention-based manner without establishing a connection (e.g., RTS/CTS in WiFi, grant procedure in LTE). Such an access pattern not only requires no orthogonal resource preallocation, which is desirable for efficient resource utilization under unpredictable traffic, but also avoids the signaling overhead and waiting time needed for connection establishment, which is desirable for the urgency constraint. However, it results in potential collisions among simultaneous packet transmissions, thus jeopardizing the network performance, especially for limited radio resources. So, it is strongly required to develop an ALOHA-like protocol that utilizes the radio resources timely and efficiently.

### B. Related Work

Considerable efforts have been made to design $p$-fixed slotted ALOHA, where the transmission probability $p$ always keeps constant, under the urgency constraint. For saturated traffic under the classical collision channel without retransmissions, [6] computed the closed-form optimal $p$ for maximizing the timely delivery ratio (TDR), defined as the percentage of packets delivered successfully before a given deadline. [7] extended this work to support an arbitrary allowed number of retransmissions using a recursive algorithm, while [8] extended this work to consider a threshold-based multiple-packet reception (MPR) channel [9] using a fixed-point iteration. For frame-synchronized traffic under the collision channel, [10] proposed a recursive algorithm to analyze the throughput, and investigated the asymptotic behavior of the optimal $p$ for maximizing the throughput. For unsynchronized periodic traffic under an SIR channel, [11] characterized the TDR in a large-scale D2D network based on the joint use of an absorbing Markov chain and the meta distribution of the SIR. Both [10],

This work was supported in part by the National Natural Science Foundation of China under Grants 62071236, 62001225, 61902256, in part by the National Science and Technology Council of Taiwan under Grant NSTC 112-2115-M-153-MY2, and in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2022D07.

Y. Zhang and J. Li are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: yijin.zhang@gmail.com; jun.li@njust.edu.cn).

A. Gong is with the School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland (e-mail: aoyu.gong@epfl.ch).

L. Deng is with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: ldeng@szu.edu.cn).

Y.-H. Lo is with the Department of Applied Mathematics, National Pingtung University, Pingtung 90003, Taiwan (e-mail: yhlo@mail.nptu.edu.tw).

Y. Lin is with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: yanlin@njust.edu.cn).

(Y. Zhang and A. Gong contributed equally to this work.)

[11] assumed no retransmission limit. For Bernoulli traffic under the collision channel, [12], [13] used stationary Markov chains to derive the optimal $p$ for maximizing the TDR under non-retransmission and non-retransmission-limit mechanisms, respectively. For a single Bernoulli-traffic user competing with saturated users under the collision channel, [7], [14] used stationary Markov chains to analyze the throughput of this user for an arbitrary allowed number of retransmissions.

To remove the restriction of fixed $p$, there has been an increasing interest in studying $p$-dynamic slotted ALOHA, where $p$ changes according to local observations, under the urgency constraint. For saturated traffic under the threshold-based MPR channel, [8] proposed to use the statistics of consecutive slots to estimate the number of active users that may suddenly changes, and then adjust $p$ to maximize the TDR based on this estimation. For frame-synchronized traffic, [15] proposed to myopically adjust $p$ for maximizing the instantaneous throughput under the collision channel when the number of active users is always known, and [16] proposed to simply double or halve $p$ based on the previous channel feedback under the threshold-based MPR channel when the number of active users is unknown; both [15], [16] applied absorbing Markov chain modeling for analysis. However, [8], [15], [16] adopted urgency-independent decision rules to adjust $p$, which may limit the performance under the urgency constraint.

There also have been many $p$-dynamic slotted ALOHA schemes without the urgency constraint. Throughput maximization is considered in [17]–[21]. [17] proposed an idea that utilizes the previous channel feedback to estimate the probability distribution of the number of active users in a Bayesian manner and then myopically adjusts $p$. A simplified implementation of this idea, called the pseudo-Bayesian algorithm, was developed in [17] for the collision channel with Poisson traffic. Under the collision channel with interrupted Poisson traffic, [18] used the statistics of consecutive idle and collision slots to accelerate the tracking process of the number of active users and then myopically adjust $p$ in a timely manner. Under the collision channel with a general traffic pattern, [19] proposed to introduce random splitting upon collisions and myopically adjust $p$ in a pseudo-Bayesian manner [17]. This work was extended by [20] to support successive interference cancellation (SIC), and by [21] to support multichannel systems. Under the collision channel with generate-at-will traffic, to improve the age of information (AoI), [22] proposed an age-dependent scheme that allows each user to transmit with a fixed probability $p$ only if its corresponding AoI exceeds a fixed threshold. Under the collision channel with Bernoulli traffic, [23] proposed a more general age-dependent scheme that allows each user to transmit with a dynamic probability $p$ if its corresponding age gain exceeds a certain threshold, which could be computed adaptively or set as a fixed value. The works using other knowledge (e.g. queuing delay, queue length) for improving AoI can be found in [24], [25]. Under the SINR-based MPR channel [9] with Poisson traffic, [26] proposed average-reward Markov Decision Process (MDP) and Partially Observable MDP (POMDP) formulations to obtain stationary optimal backlog-minimizing policies for adjusting both $p$ and transmission power in known- and unknown-backlog cases, respectively. The works using other dynamic optimization techniques (e.g. game theory, decentralized MDP) to adjust $p$ can be found in [27], [28]. Note that [17]–[21], [26]–[28] assumed that any packet can be delivered in however much time, thus the modeling approaches therein are inapplicable under the urgency constraint. Also note that [22]–[25] assumed undelivered older packets to be replaced by new packets, which causes an extremely strict "deadline" (in generate-at-will traffic) or an unpredictable "deadline" (in Bernoulli traffic), so the modeling approaches therein *cannot* be used to exactly characterize the behavior under the urgency constraint.

### C. Motivation and Contributions

In general, $p$-dynamic slotted ALOHA can be seen as a sequential access decision problem under certain observations; however, previous schemes under the urgency constraint [8], [15], [16] were designed for pursuing low complexity but without a principle of dynamic programming optimality. Instead, in this paper, we study optimal schemes to maximize the long-run performance with a focus on the frame-synchronized traffic that captures a number of IoT scenarios [1]–[3], [29], [30]. Similar to [26], we pose the access problem for different knowledge of the number of active users as an MDP and a POMDP, respectively. Although the idea of using MDPs and POMDPs in random access is not new [26], [31], our study is different because the urgency constraint plays a nontrivial role in decision making, which not only leads to defining time-dependent decision rules but also leads to answering a number of fundamental questions:

(i) How does the urgency constraint affect optimal policies?
(ii) Under the urgency constraint, under what conditions does there exist an easily implementable optimal policy, e.g., a time-independent deterministic Markovian optimal policy?
(iii) Under the urgency constraint, when it is difficult to obtain an easily implementable optimal policy, how to compute a quasi-optimal policy efficiently?

Note that the aforementioned schemes [6]–[8], [10]–[16] only take into account the deadline of the urgency constraint, but ignore the time value of successful transmissions[1].

Our key contributions lie in the following problem formulation and analysis.

(i) To characterize the time value of successful transmissions, built on a quite general non-increasing urgency function (which can be chosen according to a specific application), we generalize the traditional throughput metric to introduce a new metric: *urgency-dependent throughput* (UDT), which is defined as the long-run average expected rewards of successful transmissions per slot under the urgency constraint. The design objectives in [7], [10]–[16] can be seen as particular cases of maximizing the UDT here. In addition, to abstract channel models for the MAC layer of wireless networks, we consider a general *single-packet*

[1]A successful transmission is worth more now than a successful transmission in the future for its intended application.

*reception* (SPR) channel (i.e., a channel model where at most one packet has a chance to be successfully received when multiple packets overlap on the channel), which includes those considered in [6], [7], [10], [12], [13], [15] as particular cases.

(ii) To improve the maximum attainable UDT, we propose a novel $p$-dynamic slotted ALOHA protocol, which allows each active node to determine the current transmission probability with certainty based on not only the knowledge of the current number of active users but also the current delivery urgency. The previously known schemes for the frame-synchronized traffic [10], [15] can be seen as particular cases here.

(iii) For an idealized environment where each user always has a complete knowledge of the current number of active users, we use the theory of MDP to explicitly obtain optimal policies for maximizing the UDT, and prove that a myopic policy is in general optimal, which has been commonly believed in the literature [15], [16] as folklore knowledge[2]. We further specify a bit surprising fact[3] that the general SPR channel assumption and the non-increasing property of the urgency function are both essential to such optimality.

(iv) For a realistic environment where each user has an incomplete knowledge of the current number of active users, we use the theory of POMDP to seek optimal policies, and show that a myopic policy is in general not optimal by presenting a counterexample. Then, because of the prohibitive complexity to obtain optimal or near-optimal policies for this case, we propose a practical policy that utilizes the inherent property of our MDP framework. In addition, to reduce the activity belief updating complexity that grows linearly with the number of users, based on the reception property of the collision channel, we propose a pseudo-Bayesian belief approximation whose updating relies on two changeable parameters $M_t$, $\alpha_t$, which will be described in detail in Section VI-B, to specify a binomial distribution.

Our modeling approach can include the approaches in [10], [15], [16] as particular cases[4]. It is worth noting that, although the idealized environment is difficult to implement in practice, its performance will upper bound the maximum attainable UDT in the realistic environment, and its study will inspire the design of two practical policies for the realistic environment.

The remainder of this paper is organized as follows. The system model and related application scenarios are specified in Section II, and our access protocol is proposed in Section III. Optimal policies for the idealized and realistic environments are studied in Sections IV and V, respectively. Two practical policies for the realistic environment are presented in Section VI. Numerical results are provided in Section VII to verify our study. Section VIII draws final conclusions.

[2]Both [15], [16] believed that the myopic policy is optimal for maximizing the traditional throughput (i.e., a particular case of the UDT) under the collision channel (i.e., a particular case of the general SPR channel), but lacking a formal proof.

[3]Under the general SPR channel, a myopic policy is in general optimal means that there exists a time-independent optimal decision rule at each slot that aims to transmit packets as many as possible. Motivated by this fact, it is expected that under an arbitrary time-independent reception channel, the optimality of a myopic policy still holds.

[4]The modeling approaches in [10], [15], [16] are all based on absorbing Markov chains, which can be seen as finite-horizon MDPs with special policies.

## II. System Model and Application Scenarios

Consider a single-hop uplink system with global synchronization, consisting of a finite number, $N \geq 2$, of users and an *access point* (AP). The global time axis is based on a frame-by-frame structure, and each frame consists of $D \geq 1$ equal-duration slots. The slots in a frame are indexed from slot 1 to $D$ and the slot index set is denoted by $\mathcal{T} \triangleq \{1, 2, \ldots, D\}$. Under the frame-synchronized traffic, each user generates a single-slot packet with probability $\lambda \in (0, 1]$ at the beginning of each frame, *cannot* generate packets at other time points, and generates at most one packet per frame. Each packet is associated with a delivery deadline $D$ slots, that is, a packet generated in a frame will become useless and be removed at the end of this frame.

Each user is allowed to send a packet only at the beginning of a slot. We consider a general SPR channel, i.e., given that $0 \leq k \leq N$ packets are being transmitted in one slot, one packet is successfully received with probability $\sigma_k$, and no packet is successfully received with probability $1 - \sigma_k$, where $0 = \sigma_0 \leq \sigma_1, \sigma_2, \ldots, \sigma_N \leq 1$. Both the collision model (with or without channel errors) where $0 = \sigma_0 = \sigma_2 = \cdots = \sigma_N < \sigma_1$ and the capture model where $0 = \sigma_0 < \sigma_N < \cdots < \sigma_2 < \sigma_1$ can be seen as particular cases here. After a reception, the AP instantaneously broadcasts an *acknowledgement* (ACK) if this reception is successful, but instantaneously broadcasts a *negative ACK* (NACK) otherwise, both via an error-free control channel. Assume that both ACK and NACK transmission time are negligible compared with the slot length [10], [16].

If a packet is successfully received at slot $t$ of a frame, we assume that the AP obtains $\Gamma_t$ units of reward, where $0 < \Gamma_t \leq 1$ is a non-increasing urgency function with respect to $t$; otherwise, the AP obtains no reward. We further define the UDT as the long-run average expected reward obtained by the AP per slot[5]. The urgency function can be specified according to a given application, such as $\Gamma_t = g^{h(t-1)}$ or $t^{-h}$ with $0 < g \leq 1$ and $h \geq 0$ in online advertisement placement [33] or health monitoring [34].

At the beginning of a slot, we say a user is *active* if it has a packet to transmit; otherwise it is *inactive*. Having a complete knowledge of the values of $N$, $D$, $\lambda$, $\sigma_0, \ldots, \sigma_N$, and $\Gamma_1, \ldots, \Gamma_D$, at the beginning of each slot, each active user follows a common $p$-dynamic slotted ALOHA protocol, which will be specified in Section III, to determine stochastically whether to transmit.

Related application scenarios for frame-synchronized traffic can be found as follows.

*Industrial control*. In a periodic event-triggered control implementation [29] for closed-loop process control, multiple sensor nodes associated to a process are required to

[5]When $\Gamma_t = 1$, the UDT reduces to the conventional throughput [10], [15], and the UDT divided by $\frac{N\lambda}{D}$ reduces to the TDR [16]. When $\Gamma_t = g^{t-1}$ with some $0 < g < 1$, the UDT reduces to the throughput with discount [32].

measure the plant outputs and validate the event conditions synchronously and periodically. Then, each of the sensor nodes satisfying these conditions would send fresh measurements via a small-sized packet (usually a few bytes in length) to a machine programmable logic controller (PLC), so that the PLC can recompute the controller output and take necessary actions for the process.

*Group-based event detection*. Consider a group-based event detection application [35] where a number of sensor nodes observe the same area of interest for fault-tolerance purposes. Upon an event (or a remote request) occurrence, each node in the waking state simultaneously tries to send a report to a controller. Then the controller detects an event if it receives a certain number of positive reports from different nodes, within a certain time interval since the event occurrence.

## III. Protocol Design

In this section, we propose a $p$-dynamic slotted ALOHA protocol to specify how each active user determines its transmission probability based on the current delivery urgency and local knowledge of the number of active users. During an arbitrary frame, we denote random variable $n_t$ as the actual number of active users at the beginning of slot $t$. Clearly, $n_t \in \mathcal{N} \triangleq \{0, 1, \ldots, N\}$. We distinguish two environments for available information to obtain the transmission probability.

(i) *Idealized environment:* at the beginning of every slot $t \in \mathcal{T}$, each active user knows the actual value of $n_t$, and uses the values of $t$ and $n_t$ to compute the transmission probability $p_t$ by an access function $\widehat{\pi}_t : \mathcal{N} \to [0, 1]$. So, an access policy for this environment can be defined as $\widehat{\boldsymbol{\pi}} \triangleq [\widehat{\pi}_1, \widehat{\pi}_2, \ldots, \widehat{\pi}_D]$.

(ii) *Realistic environment:* at the beginning of every slot $t \in \mathcal{T}$, based on the available information of traffic pattern, all past AP feedback, and all past transmission probabilities, each active user $i$ obtains an activity belief, denoted by a probability vector $\mathbf{b}_{i,t} \triangleq [b_{i,t}(0), b_{i,t}(1), \ldots, b_{i,t}(N)]$ where $b_{i,t}(n)$ is the conditional probability that $n_t = n$. We also require active users to adopt the same transmission probability if they have the same activity belief at the same time. In this manner, it is obvious that $\mathbf{b}_{1,t} = \cdots = \mathbf{b}_{N,t} = \mathbf{b}_t$ since the available information becomes global. Let $\mathcal{B}_t$ denote the set of all possible values of $\mathbf{b}_t$ in $[0, 1]^{N+1}$. Then, each active user uses the values of $t$ and $\mathbf{b}_t$ to compute the transmission probability $p_t$ by an access function $\pi_t : \mathcal{B}_t \to [0, 1]$. So, an access policy for this environment can be defined as $\boldsymbol{\pi} \triangleq [\pi_1, \pi_2, \ldots, \pi_D]$.

An example to illustrate how the proposed protocol works is shown in Fig. 1.

## IV. Dynamic Optimization for the Idealized Environment

In this section, we cast the access problem for the idealized environment as an MDP, which formally leads to optimal policies, and prove that a myopic policy is in general optimal.

### A. MDP Formulation

From the access scheme specified in Section III, we see that the state process $(n_t)_{t\in\mathcal{T}}$ with the state space $\mathcal{N}$ can be viewed as a discrete-time finite-horizon, finite-state Markov chain. Now, we formulate this Markov chain $(n_t)_{t\in\mathcal{T}}$ as an MDP by describing the following definitions.

(i) *Actions:* At the beginning of each slot $t \in \mathcal{T}$, the action of each active user is its transmission probability $p_t$ taking values in the action space $[0, 1]$.

(ii) *State Transition Function:* Define the state transition function $\beta_t(n', n, p) \triangleq \Pr(n_{t+1} = n' | n_t = n, p_t = p)$. For each $t \in \mathcal{T} \setminus \{D\}$, each $n, n' \in \mathcal{N}$, and each $p \in [0, 1]$, we have

$$\beta_t(n', n, p) = \begin{cases} \eta(n, p), & \text{if } n - n' = 1, \\ 1 - \eta(n, p), & \text{if } n - n' = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\eta(n, p) \triangleq \sum_{k=0}^{n} \sigma_k \binom{n}{k} p^k (1-p)^{n-k}$.

(iii) *Reward Function:* Let $r_t(n, p)$ denote the expected reward gained at slot $t$ by the AP when $n_t = n$ and $p_t = p$. So, for each $t \in \mathcal{T}$, each $n \in \mathcal{N}$, and each $p \in [0, 1]$, we have

$$r_t(n, p) = \Gamma_t \eta(n, p). \tag{2}$$

A policy for the idealized environment defined in Section III, $\widehat{\boldsymbol{\pi}}$, can be seen as a deterministic Markovian policy here. Denote by $\widehat{\boldsymbol{\Pi}}^{\text{MD}}$ the set of all possible such polices.

Let $R^{\widehat{\boldsymbol{\pi}}}(n)$ represent the expected total reward accumulated over the time horizon $\mathcal{T}$ if the policy $\widehat{\boldsymbol{\pi}}$ is used and $n_1 = n$, which is defined by

$$R^{\widehat{\boldsymbol{\pi}}}(n) \triangleq \mathbb{E}^{\widehat{\boldsymbol{\pi}}}\Big\{ \sum_{t=1}^{D} r_t\big(n_t, \widehat{\pi}_t(n_t)\big) \mid n_1 = n \Big\}. \tag{3}$$

Averaging over all possible values of $n_1$, the UDT under the policy $\widehat{\boldsymbol{\pi}}$ can be computed by

$$\text{UDT}^{\widehat{\boldsymbol{\pi}}} = \frac{1}{D} \sum_{n\in\mathcal{N}} \binom{N}{n} \lambda^n (1-\lambda)^{N-n} R^{\widehat{\boldsymbol{\pi}}}(n). \tag{4}$$

### B. MDP Solution

Our objective is to compute

$$\widehat{\boldsymbol{\pi}}^* \in \arg\max_{\widehat{\boldsymbol{\pi}} \in \widehat{\boldsymbol{\Pi}}^{\text{MD}}} \text{UDT}^{\widehat{\boldsymbol{\pi}}}. \tag{5}$$

Since the MDP formulation in Section IV-A enjoys a finite horizon, a finite state space, a compact action space, bounded rewards, and a reward function and a state transition function that are both continuous in $p$, we obtain from [36, Prop. 4.4.3, Ch. 4] that $\widehat{\boldsymbol{\pi}}^*$ is optimal over all types of policies. Applying the backward induction algorithm [36] to the following recursive equations:

$$\begin{aligned} &U_D^*(n) = \max_{p\in[0,1]} r_D(n, p), \ \forall n \in \mathcal{N}, \\ &U_t^*(n) = \max_{p\in[0,1]} r_t(n, p) \\ &\quad + \sum_{n'\in\mathcal{N}} \beta_t(n', n, p) U_{t+1}^*(n'), \ \forall n \in \mathcal{N}, \ \forall t \in \mathcal{T} \setminus \{D\}, \end{aligned} \tag{6}$$

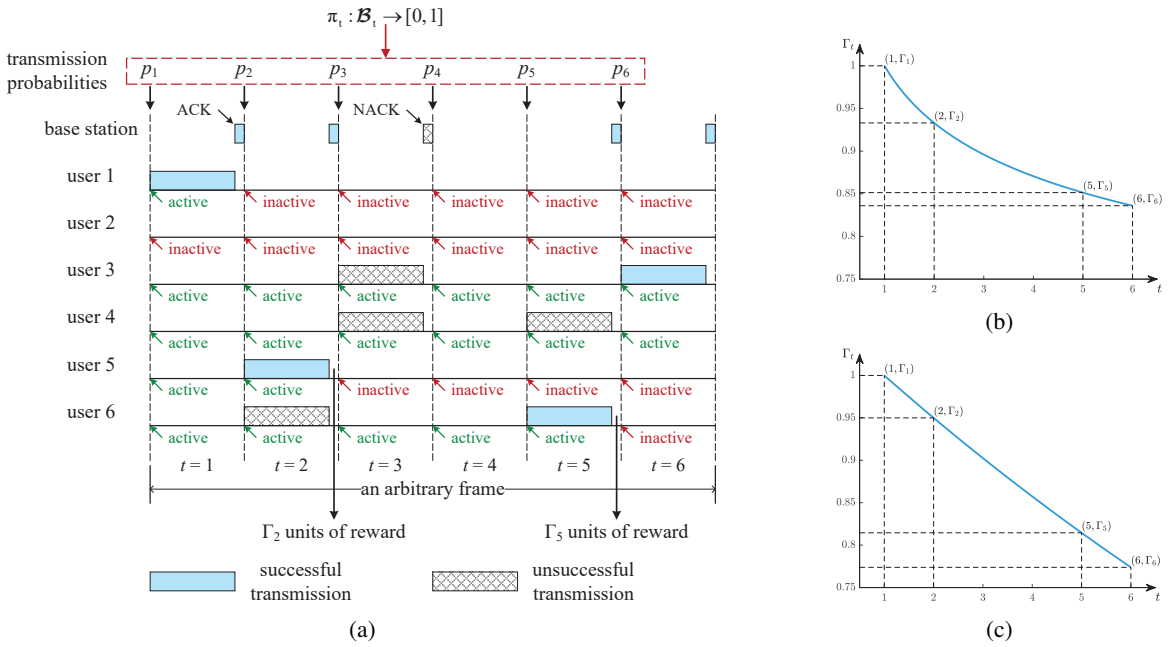



Fig. 1: An example of the working procedure of the proposed protocol for $N = 6$ and $D = 6$. (a) The working procedure. (b) $\Gamma_t = t^{-0.1}$ and UDT $= \frac{\Gamma_1+\Gamma_2+\Gamma_5+\Gamma_6}{D} = 0.6034$. (c) $\Gamma_t = 0.95^{t-1}$ and UDT $= \frac{\Gamma_1+\Gamma_2+\Gamma_5+\Gamma_6}{D} = 0.5897$.

where $U_t^*(n)$ is known as the MDP value function (defined as the maximum expected total reward from slot $t$ to $D$ when $n_t = n$), can formally lead to $\widehat{\boldsymbol{\pi}}^*$. However, it requires computing global maximizers of a number of real-coefficient univariate polynomials defined on $[0, 1]$, which is still computationally demanding.

### C. Optimality of Myopic Policy

Define a myopic policy $\widehat{\boldsymbol{\pi}}^{\text{myo}} \triangleq [\widehat{\pi}_1^{\text{myo}}, \widehat{\pi}_2^{\text{myo}}, \ldots, \widehat{\pi}_D^{\text{myo}}] \in \widehat{\boldsymbol{\Pi}}^{\text{MD}}$ that maximizes the immediate one-step reward, i.e.,

$$\widehat{\pi}_t^{\text{myo}}(n) \in \arg\max_{p\in[0,1]} r_t(n, p),\ \forall n \in \mathcal{N},\ \forall t \in \mathcal{T}. \tag{7}$$

For tractable analysis, we introduce a few more definitions which will be useful later.

Let $U_t^{\text{myo}}(n)$ denote the expected total reward from slot $t$ to $D$ for the state $n_t = n$ when each active user adopts the myopic decision rules at slots $t, t+1, \ldots, D$. So, using the finite-horizon policy evaluation algorithm [36], we have

$$\begin{aligned} &U_D^{\text{myo}}(n) = r_D\big(n, \widehat{\pi}_D^{\text{myo}}(n)\big),\ \forall n \in \mathcal{N}, \\ &U_t^{\text{myo}}(n) = r_t\big(n, \widehat{\pi}_t^{\text{myo}}(n)\big) \\ &\quad + \sum_{n'\in\mathcal{N}} \beta_t\big(n', n, \widehat{\pi}_t^{\text{myo}}(n)\big) U_{t+1}^{\text{myo}}(n'),\ \forall n \in \mathcal{N},\ \forall t \in \mathcal{T}\setminus\{D\}. \end{aligned} \tag{8}$$

By Eqs. (2), (6), (8), and $\eta^*(n) \triangleq \max_{p\in[0,1]} \eta(n, p)$, we have

$$U_t^*(0) = U_t^{\text{myo}}(0) = 0,\ \forall t \in \mathcal{T}, \tag{9}$$

$$U_D^*(n) = U_D^{\text{myo}}(n) = \Gamma_D \eta^*(n),\ \forall n \in \mathcal{N}. \tag{10}$$

Let $U_t^\diamond(n, p)$ denote the expected total reward from slot $t$ to $D$ for the state $n_t = n$ when each active user adopts the transmission probability $p_t = p$ at slot $t$ and the optimal decision rules at slots $t+1, t+2, \ldots, D$. So, we have

$$\begin{aligned} &U_D^\diamond(n, p) = r_D(n, p),\ \forall n \in \mathcal{N}, \\ &U_t^\diamond(n, p) = r_t(n, p) \\ &\quad + \sum_{n'\in\mathcal{N}} \beta_t(n', n, p) U_{t+1}^*(n'),\ \forall n \in \mathcal{N},\ \forall t \in \mathcal{T}\setminus\{D\}. \end{aligned} \tag{11}$$

We are ready for proving the optimality of $\widehat{\boldsymbol{\pi}}^{\text{myo}}$.

**Theorem 1.** *For arbitrary $N \geq 1$, $D \geq 1$, $0 \leq \sigma_1, \sigma_2, \ldots, \sigma_N \leq 1$, and non-increasing $\Gamma_t$, a myopic policy $\widehat{\boldsymbol{\pi}}^{myo}$ is optimal for the idealized environment.*

*Proof.* When $D = 1$, $\widehat{\boldsymbol{\pi}}^{\text{myo}}$ is optimal by Eqs. (6) and (8). It remains to consider the case $D > 1$.

As $U_t^*(0) = U_t^{\text{myo}}(0)$ has been settled in Eq. (9), we shall prove, for each $n \in \mathcal{N}\setminus\{0\}$, that (i) $U_t^*(n) = U_t^{\text{myo}}(n)$ by induction on $t$ from $t = D$ down to 1 and (ii) $\Gamma_{t-1} - U_t^*(n) + U_t^*(n-1) \geq 0$ by induction on $t$ from $t = D$ down to 2.

First, when $t = D$, by $0 \leq \sigma_1, \sigma_2, \ldots, \sigma_N \leq 1$, the non-increasing property of $\Gamma_t$, and Eq. (10), we have $U_D^*(n) = U_D^{\text{myo}}(n)$ and

$$\begin{aligned} &\Gamma_{D-1} - U_D^*(n) + U_D^*(n-1) \\ &\quad = \Gamma_{D-1} - \Gamma_D + \Gamma_D\big(1 - \eta^*(n) + \eta^*(n-1)\big) \geq 0, \end{aligned}$$

for each $n \in \mathcal{N}\setminus\{0\}$, thereby establishing the induction basis.

Next, when $t \in \mathcal{T}\setminus\{D\}$, we assume (i) $U_{t+1}^*(n) = U_{t+1}^{\text{myo}}(n)$ and (ii) $\Gamma_t - U_{t+1}^*(n) + U_{t+1}^*(n-1) \geq 0$ for each $n \in \mathcal{N}\setminus\{0\}$. By Eqs. (1), (2), and (11), we have

$$\begin{aligned} U_t^\diamond(n, p) &= r_t(n, p) + \sum_{n'\in\mathcal{N}} \beta_t(n', n, p) U_{t+1}^*(n') \\ &= \big(\Gamma_t - U_{t+1}^*(n) + U_{t+1}^*(n-1)\big)\eta(n, p) + U_{t+1}^*(n). \end{aligned} \tag{12}$$

Taking the derivative of Eq. (12) with respect to $p$ derives that $\frac{\mathrm{d}}{\mathrm{d}p}U_t^{\diamond}(n,p) = (\Gamma_t - U_{t+1}^*(n) + U_{t+1}^*(n-1))\frac{\mathrm{d}}{\mathrm{d}p}\eta(n,p)$. By hypothesis (ii), we obtain that $U_t^{\diamond}(n,p)$ attains its maximum when $p = \widehat{\pi}_t^{\text{myo}}(n) \in \arg\max_{p\in[0,1]} \eta(n,p)$, i.e.,

$$\begin{aligned} U_t^*(n) &= U_t^{\diamond}\big(n, \widehat{\pi}_t^{\text{myo}}(n)\big) \\ &= \big(\Gamma_t - U_{t+1}^*(n) + U_{t+1}^*(n-1)\big)\eta^*(n) + U_{t+1}^*(n). \end{aligned} \quad (13)$$

To prove $\Gamma_{t-1} - U_t^*(n) + U_t^*(n-1) \geq 0$, we further consider the following two cases.

*Case 1:* When $n = 1$, by $0 \leq \sigma_1 \leq 1$, the non-increasing property of $\Gamma_t$, and Eqs. (9), (13), we have

$$\begin{aligned} &\Gamma_{t-1} - U_t^*(1) + U_t^*(0) \\ &= \Gamma_{t-1} - \big(\Gamma_t - U_{t+1}^*(1) + U_{t+1}^*(0)\big)\eta^*(1) \\ &\quad - U_{t+1}^*(1) + U_{t+1}^*(0) \\ &= \Gamma_{t-1} - \Gamma_t + \big(1 - \eta^*(1)\big)\big(\Gamma_t - U_{t+1}^*(1)\big) \geq 0. \end{aligned}$$

*Case 2:* When $n \in \mathcal{N} \setminus \{0,1\}$, by $0 \leq \sigma_1, \sigma_2, \ldots, \sigma_n \leq 1$, the non-increasing property of $\Gamma_t$, and Eq. (13), we have

$$\begin{aligned} &\Gamma_{t-1} - U_t^*(n) + U_t^*(n-1) \\ &= \Gamma_{t-1} - \big(\Gamma_t - U_{t+1}^*(n) + U_{t+1}^*(n-1)\big)\eta^*(n) - U_{t+1}^*(n) \\ &\quad + \big(\Gamma_t - U_{t+1}^*(n-1) + U_{t+1}^*(n-2)\big) \\ &\qquad \times \eta^*(n-1) + U_{t+1}^*(n-1) \\ &= \Gamma_{t-1} - \Gamma_t + \big(1 - \eta^*(n)\big)\big(\Gamma_t - U_{t+1}^*(n) + U_{t+1}^*(n-1)\big) \\ &\quad + \eta^*(n-1)\big(\Gamma_t - U_{t+1}^*(n-1) + U_{t+1}^*(n-2)\big) \geq 0. \end{aligned}$$

In addition, by hypothesis (i) and Eqs. (7), (8), (11), (13), we further obtain

$$\begin{aligned} U_t^*(n) &= r_t\big(n, \widehat{\pi}_t^{\text{myo}}(n)\big) + \sum_{n'\in\mathcal{N}} \beta_t\big(n', n, \widehat{\pi}_t^{\text{myo}}(n)\big) U_{t+1}^*(n') \\ &= r_t\big(n, \widehat{\pi}_t^{\text{myo}}(n)\big) + \sum_{n'\in\mathcal{N}} \beta_t\big(n', n, \widehat{\pi}_t^{\text{myo}}(n)\big) U_{t+1}^{\text{myo}}(n') \\ &= U_t^{\text{myo}}(n). \end{aligned}$$

So, the inductive step is established.

Since both the base case and the inductive step have been proved as true, we have $U_t^*(n) = U_t^{\text{myo}}(n)$ for each $t \in \mathcal{T}$ and each $n \in \mathcal{N} \setminus \{0\}$. Furthermore, we obtain that

$$\begin{aligned} \text{UDT}^{\widehat{\boldsymbol{\pi}}^*} &= \frac{1}{D}\sum_{n\in\mathcal{N}} \binom{N}{n} \lambda^n (1-\lambda)^{N-n} U_1^*(n) \\ &= \frac{1}{D}\sum_{n\in\mathcal{N}} \binom{N}{n} \lambda^n (1-\lambda)^{N-n} U_1^{\text{myo}}(n) = \text{UDT}^{\widehat{\boldsymbol{\pi}}^{\text{myo}}}. \end{aligned}$$

The proof is thus complete. □

According to Theorem 1, Eqs. (2) and (7), we know that there exists an optimal transmission probability $\widehat{\pi}_t^{\text{myo}}(n)$ for each $t$, which is only dependent on the number of active users $n$ but is independent on the urgency constraint related parameters $t$, $\Gamma_t$. Further, we know this optimal transmission probability becomes smaller when $n$ increases.

### D. Discussion

We would like to emphasize that from a technical point of view, both the non-increasing property of $\Gamma_t$ and general SPR channel are the essentials to the whole proof of Theorem 1.

To this purpose, we provide an example to show that $\widehat{\boldsymbol{\pi}}^{\text{myo}}$ is in general not optimal for the idealized environment without the non-increasing property of $\Gamma_t$.

*Example 1:* Given arbitrary $N \geq 1$, $D \geq 2$, $0 < \sigma_1 \leq 1$, if $\Gamma_t < \sigma_1\Gamma_{t+1}$ for each $t \in \mathcal{T} \setminus \{D\}$, computation reveals

$$U_t^{\text{myo}}(1) < U_t^*(1), \forall t \in \mathcal{T} \setminus \{D\}, \quad (14)$$

which shows the myopic policy is not optimal. The proof of inequality (14) is in Appendix A.

We further provide an example to show that $\widehat{\boldsymbol{\pi}}^{\text{myo}}$ is in general not optimal for the idealized environment without the general SPR channel assumption.

*Example 2:* Consider a threshold-based MPR channel with capability $\gamma$, i.e., given that $1 \leq k \leq N$ packets are being transmitted in one slot, all $k$ packets will be successfully received if $1 \leq k \leq \gamma$, and no packet will be successfully received otherwise. Given arbitrary $D \geq 2$, $2 \leq \gamma < N$, and $\Gamma_t = 1$ for each $t \in \mathcal{T}$, computation reveals

$$U_t^{\text{myo}}(\gamma+1) < U_t^*(\gamma+1), \forall t \in \mathcal{T} \setminus \{D\}, \quad (15)$$

which shows the myopic policy is not optimal. The proof of inequality (15) is in Appendix B.

It remains an interesting question as to whether the results in Examples 1–2 hold for all the cases.

## V. Dynamic Optimization for the Realistic Environment

In this section, we cast the access problem for the realistic environment as a POMDP problem, show that it is infeasible to obtain optimal or near-optimal policies, and show that a myopic policy is in general not optimal.

### A. POMDP Formulation

Built on the MDP formulation specified in Section IV, we complete our POMDP formulation by describing the following definitions.

(i) *Observations and Observation Function:* At the end of each slot $t \in \mathcal{T} \setminus \{D\}$, each active user obtains an observation $o_t$ on the AP feedback, taking values from the observation space $\mathcal{O} \triangleq \{0\,(\text{no feedback received}), 1\,(\text{ACK received}), 2\,(\text{NACK received})\}$. The observation function $\omega_t(o,n,p,n') \triangleq \Pr(o_t = o | n_t = n, p_t = p, n_{t+1} = n')$ can be obtained by

$$\begin{aligned} &\omega_t(o,n,p,n') \\ &= \begin{cases} 1 - \dfrac{(1-p)^n}{1-\eta(n,p)}, & \text{if } o = 2, n = n', \eta(n,p) < 1, \\ \dfrac{(1-p)^n}{1-\eta(n,p)}, & \text{if } o = 0, n = n', \eta(n,p) < 1, \\ 1, & \text{if } o = 1, n - n' = 1, \\ 0, & \text{otherwise}, \end{cases} \end{aligned}$$

for each $t \in \mathcal{T}$, each $o \in \mathcal{O}$, each $n, n' \in \mathcal{N}$, and each $p \in [0,1]$.

(ii) *Update of the Activity Belief:* First, each active user can obtain from the traffic pattern that

$$\mathbf{b}_1 = \mathbf{h}_\lambda \triangleq \big[(1-\lambda)^N, N\lambda(1-\lambda)^{N-1}, \ldots, \lambda^N\big]. \quad (16)$$

Then, for each $t \in \mathcal{T} \setminus \{D\}$, using Bayes' rule, $\mathbf{b}_{t+1}$ can be recursively computed by

$$\begin{aligned} \mathbf{b}_{t+1} &\triangleq \theta_t(\mathbf{b}, p, o), \quad (17)\\ b_{t+1}(n') &\triangleq \Pr(n_{t+1} = n' | \mathbf{b}_t = \mathbf{b}, p_t = p, o_t = o)\\ &= \frac{\sum_{n\in\mathcal{N}} b(n)\omega_t(o,n,p,n')\beta_t(n',n,p)}{\chi_t(o,\mathbf{b},p)}, \end{aligned}$$

for each $n' \in \mathcal{N}$, where

$$\chi_t(o,\mathbf{b},p) = \sum_{n\in\mathcal{N}} b(n) \sum_{n''\in\mathcal{N}} \omega_t(o,n,p,n'')\beta_t(n'',n,p).$$

It has been shown in [37] that $\mathbf{b}_t$ is a sufficient statistic for computing optimal $p_t$ for each $t \in \mathcal{T}$.

A policy for the realistic environment defined in Section III, $\boldsymbol{\pi}$, can be seen as a deterministic Markovian policy here. Denote by $\mathbf{\Pi}^{\text{MD}}$ the set of all possible such polices.

Let $R^{\boldsymbol{\pi}}(\mathbf{h}_\lambda)$ denote the expected total reward accumulated over the time horizon $\mathcal{T}$ when $\mathbf{b}_1 = \mathbf{h}_\lambda$ and the policy $\boldsymbol{\pi}$ is employed, which is defined as

$$R^{\boldsymbol{\pi}}(\mathbf{h}_\lambda) \triangleq \mathbb{E}^{\boldsymbol{\pi}}\Big\{ \sum_{t=1}^{D} r_t\big(n_t, \pi_t(\mathbf{b}_t)\big) \mid \mathbf{b}_1 = \mathbf{h}_\lambda \Big\}.$$

Denote by $\text{UDT}^{\boldsymbol{\pi}}$ the UDT under the policy $\boldsymbol{\pi}$. We have $\text{UDT}^{\boldsymbol{\pi}} = \frac{1}{D} R^{\boldsymbol{\pi}}(\mathbf{h}_\lambda)$.

### B. POMDP Solution

Our objective is to compute

$$\boldsymbol{\pi}^* \in \arg\max_{\boldsymbol{\pi}\in\mathbf{\Pi}^{\text{MD}}} \text{UDT}^{\boldsymbol{\pi}}.$$

Since the POMDP formulation in Section IV-A enjoys a finite horizon, a finite state space, a compact action space, bounded rewards, and a reward function and $\chi_t(o,\mathbf{b},p)$ that are both continuous in $p$, we obtain from [36, Prop. 4.4.3, Ch. 4] and [38, Thm. 7.1, Ch. 6] that $\boldsymbol{\pi}^*$ is indeed optimal over all types of policies. Solving the following recursive equations:

$$\begin{aligned} V_D^*(\mathbf{b}) &= \max_{p\in[0,1]} \sum_{n\in\mathcal{N}} b(n) r_D(n,p), \ \forall \mathbf{b} \in \mathcal{B}_D,\\ V_t^*(\mathbf{b}) &= \max_{p\in[0,1]} \sum_{n\in\mathcal{N}} b(n) r_t(n,p) + \sum_{o\in\mathcal{N}} \chi_t(o,\mathbf{b},p) \quad (18)\\ &\quad \times V_{t+1}^*\big(\theta_t(\mathbf{b},p,o)\big), \ \forall \mathbf{b} \in \mathcal{B}_t, \ \forall t \in \mathcal{T}\setminus\{D\}. \end{aligned}$$

where $V_t^*(\mathbf{b})$ is known as the POMDP value function (defined as the maximum expected total reward from slot $t$ to $D$ when $\mathbf{b}_t = \mathbf{b}$), can formally lead to $\boldsymbol{\pi}^*$. However, it is computationally intractable due to the infinite belief state space $\bigcup_{t\in\mathcal{T}} \mathcal{B}_t$ and the infinite action space $[0,1]$. Even if the action space is discretized in order to compute a near-optimal policy, it is still computationally prohibitive due to super-exponential growth in the POMDP value function complexity.

### C. A counterexample for the Optimality of the Myopic Policy

In this subsection, we show that a myopic policy $\boldsymbol{\pi}^{\text{myo}} \triangleq [\pi_1^{\text{myo}}, \pi_2^{\text{myo}}, \ldots, \pi_D^{\text{myo}}] \in \mathbf{\Pi}^{\text{MD}}$ where

$$\pi_t^{\text{myo}}(\mathbf{b}) \in \arg\max_{p\in[0,1]} \sum_{n\in\mathcal{N}} b(n) r_t(n,p), \ \forall t \in \mathcal{T}, \ \forall \mathbf{b} \in \mathcal{B}_t, \quad (19)$$

is not in general optimal for the realistic environment by presenting a counterexample.

Let $V_t^{\text{myo}}(\mathbf{b})$ denote the total expected reward from slot $t$ to $D$ for $\mathbf{b}_t = \mathbf{b}$ when each active user adopts the myopic decision rules at slots $t, t+1, \ldots, D$. So, using the finite-horizon policy evaluation algorithm [37], we have

$$\begin{aligned} V_D^{\text{myo}}(\mathbf{b}) &= \sum_{n\in\mathcal{N}} b(n) r_D\big(n, \pi_D^{\text{myo}}(\mathbf{b})\big), \ \forall \mathbf{b} \in \mathcal{B}_D,\\ V_t^{\text{myo}}(\mathbf{b}) &= \sum_{n\in\mathcal{N}} b(n) r_t\big(n, \pi_t^{\text{myo}}(\mathbf{b})\big) + \sum_{o\in\mathcal{N}} \chi_t\big(o, \mathbf{b}, \pi_t^{\text{myo}}(\mathbf{b})\big)\\ &\quad \times V_{t+1}^{\text{myo}}\Big(\theta_t\big(\mathbf{b}, \pi_t^{\text{myo}}(\mathbf{b}), o\big)\Big), \ \forall \mathbf{b} \in \mathcal{B}_t, \ \forall t \in \mathcal{T}\setminus\{D\}. \end{aligned} \quad (20)$$

By Eqs. (18) and (20), we have

$$V_D^*(\mathbf{b}) = V_D^{\text{myo}}(\mathbf{b}), \ \forall \mathbf{b} \in \mathcal{B}_D. \quad (21)$$

Let $V_t^\diamond(\mathbf{b}, p)$ denote the total expected reward from slot $t$ to $D$ for $\mathbf{b}_t = \mathbf{b}$ when each active user adopts $p_t = p$ and the optimal decision rules at slots $t+1, t+2, \ldots, D$. So, we have

$$\begin{aligned} V_D^\diamond(\mathbf{b}, p) &= \sum_{n\in\mathcal{N}} b(n) r_D(n,p), \ \forall \mathbf{b} \in \mathcal{B}_D,\\ V_t^\diamond(\mathbf{b}, p) &= \sum_{n\in\mathcal{N}} b(n) r_t(n,p) + \sum_{o\in\mathcal{N}} \chi_t(o,\mathbf{b},p) \quad (22)\\ &\quad \times V_{t+1}^*\big(\theta_t(\mathbf{b},p,o)\big), \ \forall \mathbf{b} \in \mathcal{B}_t, \ \forall t \in \mathcal{T}\setminus\{D\}. \end{aligned}$$

A counterexample is shown as follows.

*Example 3:* Consider an example with $N = 2$, $D \geq 2$, $0 \leq \sigma_2 < \sigma_1 = 1$, $\Gamma_t = 1$ for each $t \in \mathcal{T}$, and $\mathbf{b}_{D-1} = [b(0), b(1), b(2)]$ satisfying $0 < b(2) \leq \frac{1}{2-2\sigma_2} b(1)$. Now we compare the following two policies: 1) $\boldsymbol{\pi}^{\text{myo}}$, and 2) $\boldsymbol{\pi}^{\text{com}}$ that adopts $p_{D-1} = \frac{1}{2-\sigma_2}$ and the myopic decision rule at slot $D$. Computation reveals that

$$V_{D-1}^{\text{myo}}(\mathbf{b}) < V_{D-1}^\diamond\Big(\mathbf{b}, \frac{1}{2-\sigma_2}\Big), \quad (23)$$

which shows that $\boldsymbol{\pi}^{\text{myo}}$ is not optimal here. The proof of Inequality (23) is provided in Appendix C. In this example, when $D = 2$, $\sigma_2 = 0.5$, and $\lambda \in (0, \frac{2}{3}]$, we can obtain

$$\frac{\text{UDT}^{\boldsymbol{\pi}^{\text{com}}} - \text{UDT}^{\boldsymbol{\pi}^{\text{myo}}}}{\text{UDT}^{\boldsymbol{\pi}^{\text{myo}}}} = \frac{V_1^\diamond(\mathbf{b}, \frac{2}{3}) - V_1^{\text{myo}}(\mathbf{b})}{V_1^{\text{myo}}(\mathbf{b})} \in (0, 0.0873],$$

showing that $\boldsymbol{\pi}^{\text{com}}$ outperforms $\boldsymbol{\pi}^{\text{myo}}$ by up to $8.73\%$ in terms of the UDT.

## VI. Two Practical Policies for the Realistic Environment

Because of the prohibitive complexity to obtain optimal or near-optimal policies from the POMDP framework, in this section, we propose a practical policy for the general SPR channel and another for the collision channel. These two policies are

both based on the QMDP approximation technique [39] owing to the following two reasons. The first is that the QMDP method approximates the POMDP value functions relying on both the MDP value functions and action-value functions (Q-functions), which allows look-ahead designs and a good complexity-benefit tradeoff. The second is that the optimality of the myopic policy in our MDP framework can be utilized to further reduce the computational complexity of MDP value functions in the QMDP method.

### *A. A Simplified QMDP-Based Policy for a general SPR channel*

The original QMDP approximation technique [39] approximates $V_t^*(\mathbf{b})$ by

$$\widehat{V}_t(\mathbf{b}) = \max_{p\in[0,1]} \sum_{n\in\mathcal{N}} b(n) Q_t^*(n,p),\ \forall \mathbf{b}\in\mathcal{B}_t,$$

where

$$\begin{aligned} &Q_D^*(n,p) = r_D(n,p),\ \forall n\in\mathcal{N},\\ &Q_t^*(n,p) = r_t(n,p)\\ &\quad + \sum_{n'\in\mathcal{N}} \beta(n',n,p) U_{t+1}^*(n'),\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T}\setminus\{D\}, \end{aligned} \tag{24}$$

are the optimal Q-functions for the MDP formulation presented in Section IV.

Although the original QMDP approximation can be used to obtain good policies efficiently, solving $U_{t+1}^*(n')$ in Eq. (24) for each $n'\in\mathcal{N}$ and each $t\in\mathcal{T}\setminus\{D\}$ is still computationally demanding in practice, as mentioned in Section IV-B. So, by Theorem 1, we replace $U_{t+1}^*(n')$ with $U_{t+1}^{\text{myo}}(n')$ in Eq. (24) to generate a simplified QMDP-based policy $\boldsymbol{\pi}^{\text{simQ}}$ for the general SPR channel with simpler and more efficient updates on $\widehat{V}_t(\mathbf{b})$. The algorithm to generate $\boldsymbol{\pi}^{\text{simQ}}$ is formally described in Algorithm 1.

**Algorithm 1** The algorithm to generate $\boldsymbol{\pi}^{\text{simQ}}$ for the general SPR channel

1: Set $t=1$. Each active user obtains $\mathbf{b}_1 = \mathbf{h}_\lambda$.
2: Given $\mathbf{b}_t = \mathbf{b}$, each active user obtains

$$\pi_t^{\text{simQ}}(\mathbf{b}) \in \arg\max_{p\in[0,1]} \sum_{n\in\mathcal{N}} b(n) Q_t^*(n,p),$$

where $Q_t^*(n,p)$ is obtained by Eq. (24) with $U_{t+1}^*(n') = U_{t+1}^{\text{myo}}(n')$ for each $n'\in\mathcal{N}$.
3: If $t\in\mathcal{T}\setminus\{D\}$, given $\mathbf{b}_t=\mathbf{b}$, $p_t = \pi_t^{\text{simQ}}(\mathbf{b})$, $o_t = o$, each active user obtains $\mathbf{b}_{t+1}$ by Eq. (17). Otherwise, stop.
4: Set $t = t+1$ and go to step 2.

### *B. A Further Simplified QMDP-Based Policy for the collision channel*

As shown in Algorithm 1, $\boldsymbol{\pi}^{\text{simQ}}$ for the general SPR channel requires the full Bayesian updating of the activity belief $\mathbf{b}_t$. In this subsection, based on $\boldsymbol{\pi}^{\text{simQ}}$, we propose a further simplified QMDP-based policy $\boldsymbol{\pi}^{\text{furQ}}$ for the collision channel, which allows each active user to update $\mathbf{b}_t$ (in a pseudo-Bayesian manner) more efficiently relying on the special reception property of such a channel.

Denote by $\mathbf{b}_t^{\text{bd}} \triangleq \left[b_t^{\text{bd}}(0), b_t^{\text{bd}}(1), \ldots, b_t^{\text{bd}}(N)\right]$ an approximation of $\mathbf{b}_t$. It is specified by a binomial distribution with a parameter vector $(M_t,\alpha_t)$, i.e., when $(M_t,\alpha_t)=(M,\alpha)$,

$$b_t^{\text{bd}}(n) = \begin{cases} \binom{M}{n}\alpha^n(1-\alpha)^{M-n}, & \text{if } 0\le n\le M,\\ 0, & \text{otherwise.}\end{cases} \tag{25}$$

In such a way, each active user only needs to maintain the parameter vector $(M_t,\alpha_t)$.

By Eq. (16), we have $(M_1,\alpha_1)=(N,\lambda)$ and $\mathbf{b}_1 = \mathbf{b}_1^{\text{bd}}$. For each $t\in\mathcal{T}\setminus\{D\}$, given $(M_t,\alpha_t)=(M,\alpha)$, $\mathbf{b}_t^{\text{bd}} = \mathbf{b}^{\text{bd}}$, and $p_t = p$, we first compute an intermediate variable $\mathbf{b}_{t+1}^{\text{med}} \triangleq \left[b_{t+1}^{\text{med}}(0), b_{t+1}^{\text{med}}(1), \ldots, b_{t+1}^{\text{med}}(N)\right]$ and then update $(M_{t+1},\alpha_{t+1})$. Three cases for different observations are considered as follows.

*The case $o_t=0$:* By the Bayes' rule, we obtain

$$\begin{aligned} b_{t+1}^{\text{med}}(n') &= \frac{\sum_{n\in\mathcal{N}} b^{\text{bd}}(n)\omega_t(0,n,p,n')\beta_t(n',n,p)}{\chi_t(0,\mathbf{b}^{\text{bd}},p)}\\ &= \begin{cases} \binom{M}{n'}\left(\frac{\alpha-\alpha p}{1-\alpha p}\right)^{n'}\left(1-\frac{\alpha-\alpha p}{1-\alpha p}\right)^{M-n'}, & \text{if } 0\le n'\le M,\\ 0, & \text{otherwise,}\end{cases} \end{aligned}$$

which follows the form of Eq. (25). So, we require $\mathbf{b}_{t+1}^{\text{bd}}$ to directly take the value of $\mathbf{b}_{t+1}^{\text{med}}$ and set

$$(M_{t+1},\alpha_{t+1}) = \Big(M, \frac{\alpha-\alpha p}{1-\alpha p}\Big). \tag{26}$$

*The case $o_t=1$:* When $M>1$, we have $\alpha p<1$. By the Bayes' rule, we obtain

$$\begin{aligned} b_{t+1}^{\text{med}}(n') &= \frac{\sum_{n\in\mathcal{N}} b^{\text{bd}}(n)\omega_t(1,n,p,n')\beta_t(n',n,p)}{\chi_t(1,\mathbf{b}^{\text{bd}},p)}\\ &= \begin{cases} \binom{M-1}{n'}\left(\frac{\alpha-\alpha p}{1-\alpha p}\right)^{n'}\left(1-\frac{\alpha-\alpha p}{1-\alpha p}\right)^{M-1-n'}, & \text{if } 0\le n'< M,\\ 0, & \text{otherwise,}\end{cases} \end{aligned}$$

which still follows the form of Eq. (25). We then require $\mathbf{b}_{t+1}^{\text{bd}}$ to again directly take the value of $\mathbf{b}_{t+1}^{\text{med}}$ and set

$$(M_{t+1},\alpha_{t+1}) = \Big(M-1, \frac{\alpha-\alpha p}{1-\alpha p}\Big). \tag{27}$$

When $M=1$, it is obvious that

$$(M_{t+1},\alpha_{t+1}) = (0,0). \tag{28}$$

*The case $o_t=2$:* When $\alpha p<1$, by the Bayes' rule, we obtain

$$\begin{aligned} b_{t+1}^{\text{med}}(n') &= \frac{\sum_{n\in\mathcal{N}} b^{\text{bd}}(n)\omega_t(2,n,p,n')\beta_t(n',n,p)}{\chi_t(2,\mathbf{b}^{\text{bd}},p)}\\ &= \Bigg[\binom{M}{n'}\big(\alpha(1-p)\big)^{n'}\big(1-\alpha(1-p)\big)^{M-n'}\\ &\quad -(1-\alpha p)^M\binom{M}{n'}\Big(\frac{\alpha-\alpha p}{1-\alpha p}\Big)^{n'}\Big(1-\frac{\alpha-\alpha p}{1-\alpha p}\Big)^{M-n'}\\ &\quad -\sigma M\alpha p(1-\alpha p)^{M-1}\\ &\qquad \times\binom{M-1}{n'}\Big(\frac{\alpha-\alpha p}{1-\alpha p}\Big)^{n'}\Big(1-\frac{\alpha-\alpha p}{1-\alpha p}\Big)^{M-1-n'}\Bigg]\\ &\quad \times\big(1-(1-\alpha p)^M-\sigma M\alpha p(1-\alpha p)^{M-1}\big)^{-1}, \end{aligned}$$

for $n' = 0, 1, \ldots, M$, which does not follow the form of Eq. (25). For the sake of consistency, under the premise of unchanged mean, we modify the value of $\mathbf{b}_{t+1}^{\text{med}}$ to follow the form of Eq. (25), and set

$$(M_{t+1}, \alpha_{t+1}) = \Big( M, \big(\alpha - (\alpha - \alpha p)(1-\alpha p)^{M-1} - \sigma \alpha p (1-\alpha p)^{M-2}(M\alpha - M\alpha p - \alpha + 1)\big) \times \big(1 - (1-\alpha p)^M - \sigma M \alpha p (1-\alpha p)^{M-1}\big)^{-1} \Big). \quad (29)$$

When $\alpha p = 1$, it is obvious that

$$(M_{t+1}, \alpha_{t+1}) = (M, 1). \quad (30)$$

Based on the above belief approximation, the algorithm to generate $\boldsymbol{\pi}^{\text{furQ}}$ is formally described in Algorithm 2.

**Algorithm 2** The algorithm to generate $\boldsymbol{\pi}^{\text{furQ}}$ for the collision channel

1: Set $t = 1$. Each active user obtains $(M_1, \alpha_1) = (N, \lambda)$.
2: Given $(M_t, \alpha_t) = (M, \alpha)$, each active user obtains
$$\pi_t^{\text{furQ}}(M, \alpha) \in \arg\max_{p \in [0,1]} \sum_{n=0}^{M} \binom{M}{n} \alpha^n (1-\alpha)^{M-n} Q_t^*(n, p).$$
where $Q_t^*(n,p)$ is obtained by Eq. (24) with $U_{t+1}^*(n') = U_{t+1}^{\text{myo}}(n')$ for each $n' \in \mathcal{N}$.
3: If $t \in \mathcal{T} \setminus \{D\}$, given $(M_t, \alpha_t) = (M, \alpha)$, $p_t = \pi_t^{\text{furQ}}(M, \alpha)$, $o_t = o$, each active user obtains $(M_{t+1}, \alpha_{t+1})$ by Eqs. (26)–(30) accordingly. Otherwise, stop.
4: Set $t = t + 1$ and go to step 2.

Evaluating the belief approximation error at slot $t$ under $\boldsymbol{\pi}^{\text{furQ}}$ is a complicated issue since it is dependent on all previous channel observations before slot $t$. Obviously, if no NACK is observed before slot $t$, there is no error at slot $t$. But as soon as a NACK is observed at slot $t'$ ($t' < t$), each update at slot $t', t'+1, \ldots, t-1$ would yield a complicated (positive or negative) impact on the error at slot $t$, which is difficult to be quantified. Roughly speaking, as $t$ increases, the error fluctuation at slot $t$ would increase, so the approximation effectiveness at slot $t$ would decrease. In Table I, we provide realizations of $\mathbf{b}_t$ and its approximation $\mathbf{b}_t^{\text{bd}}$ when $\boldsymbol{\pi}^{\text{furQ}}$ is used for $N = 8$, $\lambda = 0.8$, $D = 20$, $\sigma_1 = 0.95$, $\sigma_0 = \sigma_2 = \cdots = \sigma_N = 0$. The Bhattacharyya distance between $\mathbf{b}_t$ and $\mathbf{b}_t^{\text{bd}}$ is at most $0.011738$, showing that the approximation method is effective. Furthermore, as $D$ usually takes a small value in real-time services [1]–[3], it is expected that our method would enjoy acceptable effectiveness, which will be examined in Section VII.

*Remark:* Under $\boldsymbol{\pi}^{\text{simQ}}$, the computational complexity of matrix multiplication to obtain $\mathbf{b}_t$ at the beginning of slot $t$ is $O(|\mathcal{N}|^2)$, the same as under $\boldsymbol{\pi}^*$. Under $\boldsymbol{\pi}^{\text{furQ}}$, such complexity is reduced to $O(1)$ owing to the proposed belief approximation. To generate $p_t$ at the beginning of slot $t$ under $\boldsymbol{\pi}^*$, we need to first obtain all possible $\mathbf{b}_{t'}$ in $\mathcal{B}_{t'}$ for $t' = t, t+1, \ldots, D$, and then find global maximizers of $\sum_{t'=t}^{D} |\mathcal{B}_{t'}|$ real coefficient univariate polynomials. To generate $p_t$ at the beginning of slot $t$ under $\boldsymbol{\pi}^{\text{simQ}}$ or $\boldsymbol{\pi}^{\text{furQ}}$, we only need to find a global maximizer of a polynomial after obtaining $\mathbf{b}_t$ and the optimal Q-functions (which are derived by finding global maximizers of $|\mathcal{N}|$ polynomials), owing to the look-ahead designs and Theorem 1. So, $\boldsymbol{\pi}^{\text{simQ}}$ and $\boldsymbol{\pi}^{\text{furQ}}$ both enjoy significantly lower complexity than $\boldsymbol{\pi}^*$.

## VII. Numerical Results

To validate the studies in Sections IV–VI, this section compares the UDT performance of the optimal policy for the idealized environment $\widehat{\boldsymbol{\pi}}^*$, the simplified QMDP-based policy for the realistic environment $\boldsymbol{\pi}^{\text{simQ}}$, the further simplified QMDP-based policy for the realistic environment $\boldsymbol{\pi}^{\text{furQ}}$ (only applicable to the collision channel), the myopic policy for the realistic environment $\boldsymbol{\pi}^{\text{myo}}$, the optimal static scheme[6] [10] for the realistic environment $\boldsymbol{\pi}^{\text{sta}}$, and the D&H scheme[7] [16] for the realistic environment. $\Gamma_t = t^{-0.1}$ and $\Gamma_t = 0.95^{t-1}$ are both considered[8]. This section also compares the *packet loss ratio* (PLR) of the six schemes under sporadic traffic, which is a primary concern of high-reliability IoT [2]. Minimizing the PLR here is identical to maximizing the UDT with $\Gamma_t = 1$.

In the first two subsections, we comply with the system model specified in Section II to set up the numerical experiments, and shall change the network configuration over a broad range to examine the impact of access design on the UDT and PLR. In the third subsection, we relax some system assumptions to examine the robustness of the proposed policies. Each result is an average in 10 independent numerical experiments, each of which lasts for $10^6$ frames, running in MATLAB over an 8-core AMD Ryzen 7 5800H 3.2GHz CPU and 16GB memory.

### A. Performance: Collision channel

Consider the collision channel with $\sigma_1 = 0.95$. Figs. 2(a)–(c) show the UDT as a function of the channel load $N\lambda/D$ for $N = 50$ and $\Gamma_t = t^{-0.1}$. We observe that $\boldsymbol{\pi}^{\text{simQ}}$ always performs best in the realistic environment, that is, enjoys $0.22\%$ – $12.01\%$ improvement over $\boldsymbol{\pi}^{\text{myo}}$, $1.85\%$ – $10.78\%$ improvement over $\boldsymbol{\pi}^{\text{sta}}$, and $2.15\%$ – $51.45\%$ improvement over $\boldsymbol{\pi}^{\text{dah}}$. The reason is obvious that $\boldsymbol{\pi}^{\text{sta}}$ utilizes no feedback information to optimize its access pattern, $\boldsymbol{\pi}^{\text{dah}}$ utilizes the feedback information in an unrigorous manner, and $\boldsymbol{\pi}^{\text{myo}}$ is only one-step look-ahead. This also conforms that $\boldsymbol{\pi}^{\text{myo}}$ is not in general optimal as shown in Section IV-C. Meanwhile,

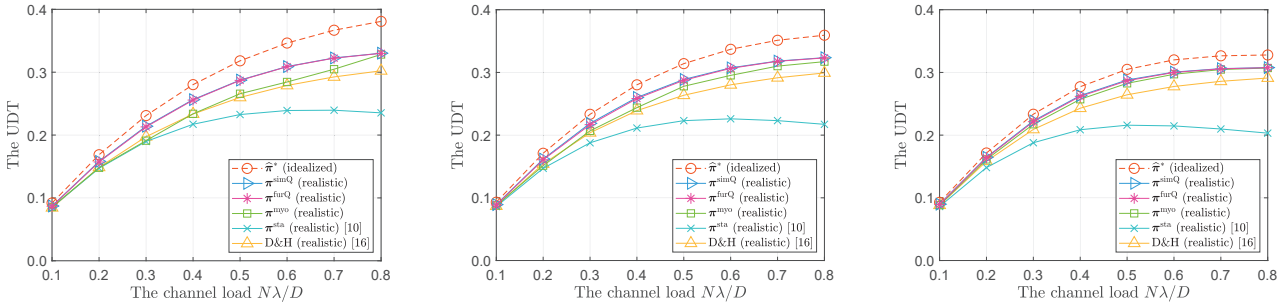
[6] A particular policy in $\boldsymbol{\Pi}^{\text{MD}}$ that allows each active user to adopt an optimal fixed and identical transmission probability.

[7] A scheme that allows each active user to adopt the transmission probability $p_1 = 1$ in the first slot of a frame, $p_{t+1} = p_t$ if $o_t = 0$, $p_{t+1} = \min(2p_t, 1)$ if $o_t = 1$, and $p_{t+1} = p_t/2$ if $o_t = 2$.

[8] The choice of the urgency function does not effect any theoretical results or algorithms in this paper, but might effect the preference for transmission probabilities. To characterize the desire for data refreshing in online advertisement placement and online Web ranking [33], we employ $\Gamma_t = t^{-0.1}$ that decreases more slowly when $t$ becomes larger, since the depreciation speed of data in these applications becomes lower with the passing of time. To account for the time value of rewards, discounting arises naturally in many applications of stochastic dynamic programming, such as ecology, economics, and communications engineering [36]. As such, we select $\Gamma_t = 0.95^{t-1}$ to reflect the impact of discounting.

TABLE I: Realizations of $\mathbf{b}_t$ and its approximation $\mathbf{b}_t^{\text{bd}}$ when $\boldsymbol{\pi}^{\text{furQ}}$ is used for $N = 8$, $\lambda = 0.8$, $D = 20$, $\sigma_1 = 0.95$, $\sigma_0 = \sigma_2 = \cdots = \sigma_N = 0$, and their Bhattacharyya distance $Bd$.

| | | $b_t(0)$ | $b_t(1)$ | $b_t(2)$ | $b_t(3)$ | $b_t(4)$ | $b_t(5)$ | $b_t(6)$ | $b_t(7)$ | $b_t(8)$ | $Bd$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $t=1$ | $\mathbf{b}_t$ | 0.000003 | 0.000082 | 0.001147 | 0.009175 | 0.045875 | 0.146801 | 0.293601 | 0.335544 | 0.167772 | 0 |
| $o_1=1$ | Approx. | 0.000003 | 0.000082 | 0.001147 | 0.009175 | 0.045875 | 0.146801 | 0.293601 | 0.335544 | 0.167772 | |
| $t=2$ | $\mathbf{b}_t$ | 0.000029 | 0.000706 | 0.007288 | 0.041816 | 0.143945 | 0.297307 | 0.341145 | 0.167764 | 0 | 0 |
| $o_2=0$ | Approx. | 0.000029 | 0.000706 | 0.007288 | 0.041816 | 0.143945 | 0.297307 | 0.341145 | 0.167764 | 0 | |
| $t=3$ | $\mathbf{b}_t$ | 0.000073 | 0.001486 | 0.012916 | 0.062357 | 0.180632 | 0.313947 | 0.303142 | 0.125446 | 0 | 0 |
| $o_3=0$ | Approx. | 0.000073 | 0.001486 | 0.012916 | 0.062357 | 0.180632 | 0.313947 | 0.303142 | 0.125446 | 0 | |
| $t=4$ | $\mathbf{b}_t$ | 0.000180 | 0.003059 | 0.022273 | 0.090108 | 0.218722 | 0.318546 | 0.257738 | 0.089374 | 0 | 0 |
| $o_4=0$ | Approx. | 0.000180 | 0.003059 | 0.022273 | 0.090108 | 0.218722 | 0.318546 | 0.257738 | 0.089374 | 0 | |
| $t=5$ | $\mathbf{b}_t$ | 0.000433 | 0.006133 | 0.037230 | 0.125553 | 0.254050 | 0.308434 | 0.208033 | 0.060135 | 0 | 0 |
| $o_5=2$ | Approx. | 0.000433 | 0.006133 | 0.037230 | 0.125553 | 0.254050 | 0.308434 | 0.208033 | 0.060135 | 0 | |
| $t=6$ | $\mathbf{b}_t$ | 0 | 0.000254 | 0.007856 | 0.058187 | 0.194429 | 0.336873 | 0.296784 | 0.105617 | 0 | 0.001350 |
| $o_6=0$ | Approx. | 0.000083 | 0.001637 | 0.013902 | 0.065587 | 0.185655 | 0.315314 | 0.297514 | 0.120308 | 0 | |
| $t=7$ | $\mathbf{b}_t$ | 0 | 0.000533 | 0.013728 | 0.084740 | 0.235966 | 0.340708 | 0.250141 | 0.074183 | 0 | 0.002052 |
| $o_7=0$ | Approx. | 0.000207 | 0.003423 | 0.024224 | 0.095238 | 0.224659 | 0.317971 | 0.250023 | 0.084255 | 0 | |
| $t=8$ | $\mathbf{b}_t$ | 0 | 0.001100 | 0.023438 | 0.119758 | 0.276044 | 0.329929 | 0.200508 | 0.049223 | 0 | 0.003384 |
| $o_8=2$ | Approx. | 0.000511 | 0.006986 | 0.040925 | 0.133186 | 0.260066 | 0.304689 | 0.198316 | 0.055320 | 0 | |
| $t=9$ | $\mathbf{b}_t$ | 0 | 0.000044 | 0.004952 | 0.055623 | 0.211188 | 0.358942 | 0.283817 | 0.085434 | 0 | 0.004467 |
| $o_9=1$ | Approx. | 0.000099 | 0.001888 | 0.015494 | 0.070624 | 0.193147 | 0.316938 | 0.288928 | 0.112883 | 0 | |
| $t=10$ | $\mathbf{b}_t$ | 0.000019 | 0.003581 | 0.049883 | 0.208761 | 0.366654 | 0.287604 | 0.083498 | 0 | 0 | 0.003579 |
| $o_{10}=0$ | Approx. | 0.000832 | 0.011283 | 0.063773 | 0.192244 | 0.325981 | 0.294802 | 0.111086 | 0 | 0 | |
| $t=15$ | $\mathbf{b}_t$ | 0.000017 | 0.008103 | 0.124040 | 0.386871 | 0.373773 | 0.107195 | 0 | 0 | 0 | 0.011738 |
| $o_{15}=1$ | Approx. | 0.001804 | 0.022897 | 0.116213 | 0.294923 | 0.374224 | 0.189939 | 0 | 0 | 0 | |
| $t=20$ | $\mathbf{b}_t$ | 0.776553 | 0.212644 | 0.010803 | 0 | 0 | 0 | 0 | 0 | 0 | 0.002857 |
| $o_{20}=0$ | Approx. | 0.718406 | 0.258365 | 0.023229 | 0 | 0 | 0 | 0 | 0 | 0 | |




(a) $\Gamma_t = t^{-0.1}$, $D = 3$. (b) $\Gamma_t = t^{-0.1}$, $D = 5$. (c) $\Gamma_t = t^{-0.1}$, $D = 10$.

(d) $\Gamma_t = 0.95^{t-1}$, $D = 3$. (e) $\Gamma_t = 0.95^{t-1}$, $D = 5$. (f) $\Gamma_t = 0.95^{t-1}$, $D = 10$.

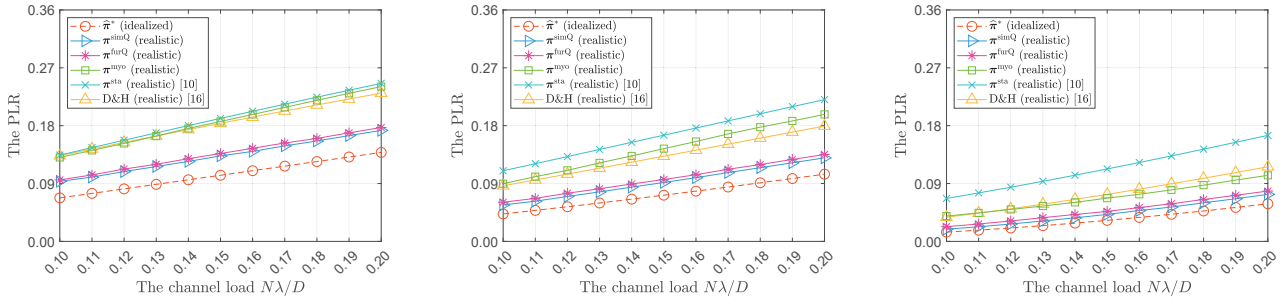
Fig. 2: The UDT as a function of the channel load $N\lambda/D$ for $N = 50$ under the collision channel with $\sigma_1 = 0.95$.

we observe that $\boldsymbol{\pi}^{\text{furQ}}$ performs very close to $\boldsymbol{\pi}^{\text{simQ}}$ with only $0.04\%$ – $0.99\%$ loss, which conforms that the belief approximation in Section VI-B is reasonable. We note that, as expected, such loss increases with $D$ since more approximation errors would be introduced, but is always small since the time value of successful transmissions would weaken the negative impact of approximation errors for large $D$. In addition, we observe that the gap between the UDT of $\widehat{\boldsymbol{\pi}}^*$ and $\boldsymbol{\pi}^{\text{simQ}}$ is minor when $N\lambda/D$ is small and gradually becomes more noticeable as $N\lambda/D$ increases. This phenomenon is due to the fact: when $\lambda < 0.5$, the variance of $n_1$ becomes larger as $\lambda$ increases, so the impact of complete knowledge of $n_t$ becomes more significant.

Figs. 2(d)–(f)[9] compare the UDT performance of the six schemes when $N = 50$ and $\Gamma_t = 0.95^{t-1}$. We observe that $\boldsymbol{\pi}^{\text{simQ}}$ enjoys $0.21\%$ – $12.30\%$ improvement over $\boldsymbol{\pi}^{\text{myo}}$, $2.40\%$ – $10.77\%$ improvement over $\boldsymbol{\pi}^{\text{sta}}$, and $2.66\%$ – $55.37\%$ improvement over $\boldsymbol{\pi}^{\text{dah}}$, and observe that $\boldsymbol{\pi}^{\text{furQ}}$ performs very close to $\boldsymbol{\pi}^{\text{simQ}}$ with only $0.04\%$ – $0.73\%$ loss. The results clearly verify again the benefits of utilizing the feedback information in sequential decision making of $\boldsymbol{\pi}^{\text{simQ}}$ and $\boldsymbol{\pi}^{\text{furQ}}$.

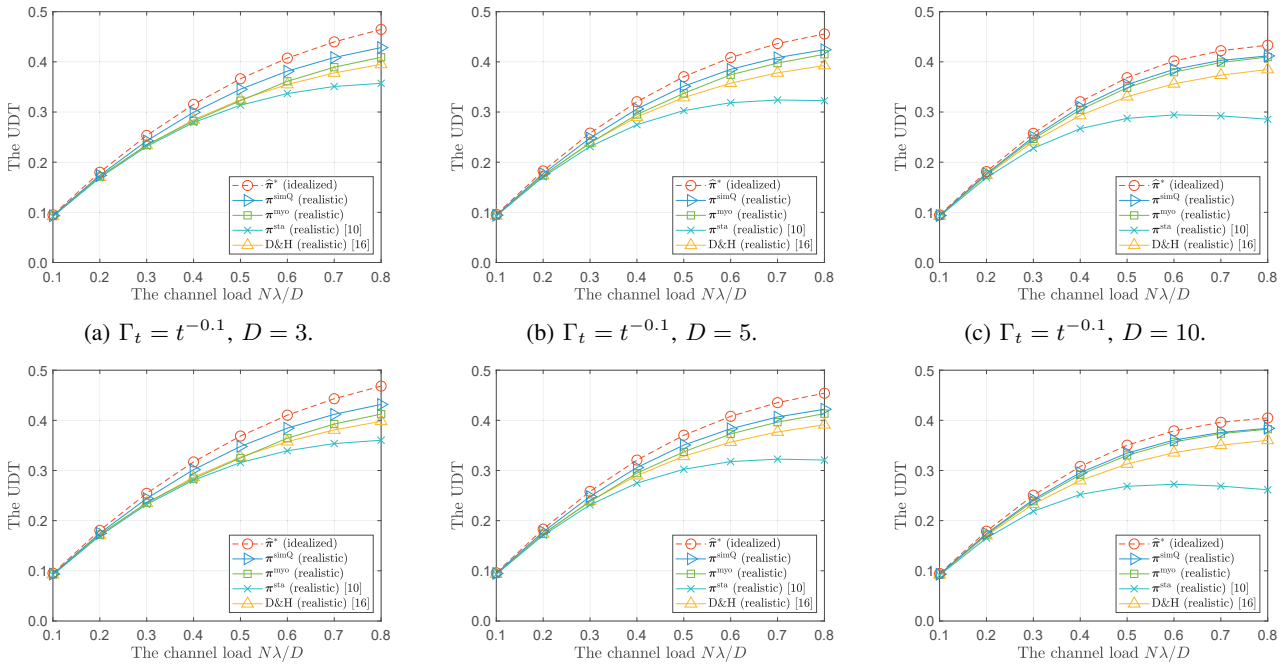
Fig. 3 compares the PLR of the six schemes under sporadic

[9]As $\sum_{t=1}^{3} t^{-0.1}$ and $\sum_{t=1}^{3} 0.95^{t-1}$ take almost the same value, Figs. 2(a)(d) for $D = 3$ look almost the same. Similar observations can be found in Figs. 2(b)(e) for $D = 5$. On the contrary, as $\sum_{t=1}^{10} t^{-0.1}$ is obviously larger than $\sum_{t=1}^{10} 0.95^{t-1}$, the UDT in Fig. 2(c) is obviously larger than that in Fig. 2(f) for $D = 10$.




(a) $D = 3$. (b) $D = 5$. (c) $D = 10$.

Fig. 3: The PLR as a function of the channel load $N\lambda/D$ for $N = 100$ under the collision channel with $\sigma_1 = 0.95$.




(a) $\Gamma_t = t^{-0.1}$, $D = 3$. (b) $\Gamma_t = t^{-0.1}$, $D = 5$. (c) $\Gamma_t = t^{-0.1}$, $D = 10$.

(d) $\Gamma_t = 0.95^{t-1}$, $D = 3$. (e) $\Gamma_t = 0.95^{t-1}$, $D = 5$. (f) $\Gamma_t = 0.95^{t-1}$, $D = 10$.

Fig. 4: The UDT as a function of the channel load $N\lambda/D$ for $N = 50$ under the SINR-based capture channel with $\nu = 1$, $\kappa = 20$.

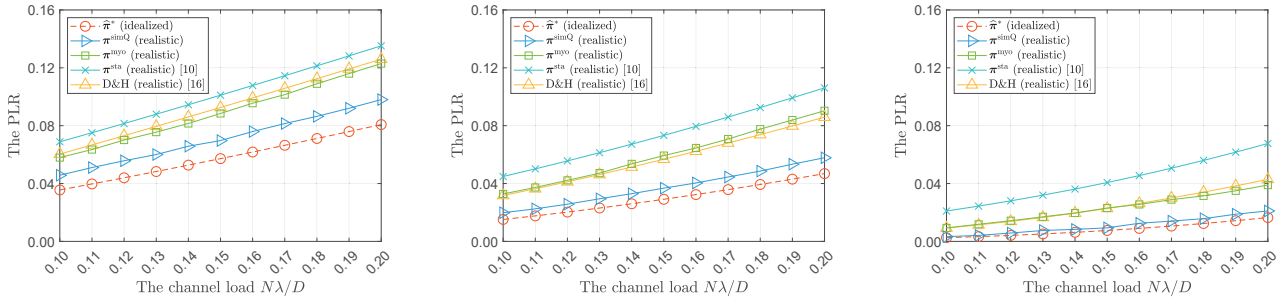
traffic generated from $N = 100$ users, which is equivalent to comparing the UDT with $\Gamma_t = 1$. We observe that $\boldsymbol{\pi}^{\text{simQ}}$ enjoys $28.31\%$ – $51.86\%$ PLR reduction over $\boldsymbol{\pi}^{\text{myo}}$, $25.15\%$ – $49.54\%$ reduction over $\boldsymbol{\pi}^{\text{sta}}$, and $29.78\%$ – $71.61\%$ reduction over $\boldsymbol{\pi}^{\text{dah}}$. We also observe that $\boldsymbol{\pi}^{\text{furQ}}$ still performs close to $\boldsymbol{\pi}^{\text{simQ}}$ in terms of PLR. The results indicate that both $\boldsymbol{\pi}^{\text{simQ}}$ and $\boldsymbol{\pi}^{\text{furQ}}$ are suitable random access candidates in scenarios that focus on the reliability target.

### *B. Performance: SINR-based Capture*

Consider an SINR-based capture model, where $\sigma_n$ can be computed by $\sigma_n = \frac{e^{-\frac{\nu}{\kappa}}}{(1+\nu)^{n-1}}$, $\forall n \in \mathcal{N} \setminus \{0\}$, and refer our readers to [40] for more details on the model. Here, we set $\nu = 1$, $\kappa = 20$. The UDT advantage of $\boldsymbol{\pi}^{\text{simQ}}$ is confirmed again in Fig. 4, which shows the UDT as a function of the channel load $N\lambda/D$ for $N = 50$ under the capture model. We observe that $\boldsymbol{\pi}^{\text{simQ}}$ enjoys $0.20\%$ – $7.11\%$ improvement over $\boldsymbol{\pi}^{\text{myo}}$, $1.19\%$ – $8.45\%$ improvement over $\boldsymbol{\pi}^{\text{sta}}$, $1.44\%$ – $46.84\%$ improvement over $\boldsymbol{\pi}^{\text{dah}}$. It is interesting to note that the gap between the UDT of $\widehat{\boldsymbol{\pi}}^*$ and $\boldsymbol{\pi}^{\text{simQ}}$ is less notable than that in Fig. 2, which can be attribute to the fact that increasing the reception capability weakens the benefit of complete knowledge of $n_t$. The significant PLR advantage of $\boldsymbol{\pi}^{\text{simQ}}$ is confirmed under the capture model in Fig. 5, which shows that $\boldsymbol{\pi}^{\text{simQ}}$ enjoys $19.29\%$ – $65.01\%$ PLR reduction over $\boldsymbol{\pi}^{\text{myo}}$, $22.17\%$ – $64.48\%$ reduction over $\boldsymbol{\pi}^{\text{sta}}$, and $27.49\%$ – $84.45\%$ reduction over $\boldsymbol{\pi}^{\text{dah}}$.

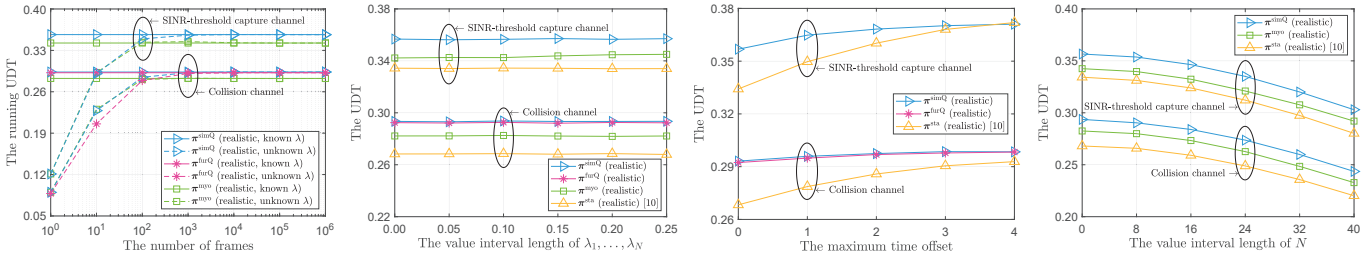
### *C. Robustness to Relaxed System Assumptions*

When the packet generation probability $\lambda$ is unavailable, based on local packet arrival events, each user can locally estimate the value of $\lambda$ but at the cost of low convergence




(a) $D = 3$. (b) $D = 5$. (c) $D = 10$.

Fig. 5: The PLR as a function of the channel load $N\lambda/D$ for $N = 100$ under the SINR-based capture channel with $\nu = 1$, $\kappa = 20$.




(a) Unknown $\lambda$. (b) Heterogeneous $\lambda_1, \ldots, \lambda_N$. (c) Unsynchronized traffic. (d) Dynamic-varying $N$.

Fig. 6: The UDT under relaxed system assumptions for $\lambda = 0.125$, $\Gamma_t = t^{-0.1}$, $D = 5$ under the collision channel with $\sigma_1 = 0.95$ and the SINR-based capture channel with $\nu = 1$, $\kappa = 20$. $N$ is fixed to 20 in (a)–(c) and is varying in (d).

speed (due to the sporadic traffic assumption)[10]. Then, each active user uses the estimated value of $\lambda$ to obtain the initial belief in the POMDP formulation, and compute the policies. Since different active users may have different estimated values of $\lambda$ and adopt different transmission probabilities at the same slot, as shown in Fig. 6(a) for a single numerical experiment, the system performance converges to a stable value within a certain number of frames.

Under different packet generation probabilities $\lambda_1, \ldots, \lambda_N$, for the idealized environment, there is no impact on the MDP formulation and the optimality of the myopic policy, but the UDT needs to be recalculated as $\frac{1}{D}\sum_{n\in\mathcal{N}}\sum_{\mathcal{I}\in\mathbb{P}(\mathcal{N}),|\mathcal{I}|=n}\prod_{i\in\mathcal{I}}\lambda_i\prod_{j\in\mathcal{N}\setminus\mathcal{I}}(1-\lambda_j)R^{\widehat{\boldsymbol{\pi}}}(n)$, where $\mathbb{P}(\mathcal{N})$ is the power set of $\mathcal{N}$. For the realistic environment, such a heterogeneous scenario would require us to set the initial belief as $[\prod_{n\in\mathcal{N}}(1-\lambda_n), \sum_{n\in\mathcal{N}}\lambda_n\prod_{n'\in\mathcal{N},n'\neq n}(1-\lambda_{n'}), \ldots, \prod_{n\in\mathcal{N}}\lambda_n]$, and make the policy $\boldsymbol{\pi}^{\text{furQ}}$ no longer applicable as $\boldsymbol{\pi}^{\text{furQ}}$ requires the initial belief to be a Binomial distribution. Assuming $\lambda_1, \lambda_2, \ldots, \lambda_N$ are uniformly and randomly chosen from an interval with the mean $\lambda$, Fig. 6(b) shows that $\boldsymbol{\pi}^{\text{simQ}}$ outperforms other policies for different interval lengths: 3.85% – 4.08% improvement over $\boldsymbol{\pi}^{\text{myo}}$ under the collision channel, 9.21% – 9.53% improvement over $\boldsymbol{\pi}^{\text{sta}}$ under the collision channel, 3.44% – 4.25% improvement over $\boldsymbol{\pi}^{\text{myo}}$ under the SINR-based capture channel, and 6.63% – 6.90% over $\boldsymbol{\pi}^{\text{sta}}$ under the SINR-based capture channel.

Consider that the first frames of different users may have time offsets taking arbitrary values uniformly and randomly from $\{0, 1, \ldots, \Delta\}$ with $0 \leq \Delta \leq D-1$. Obviously, the theoretical work in this paper is inapplicable to such unsynchronized periodic traffic[11]. Fig. 6(c) shows the robustness of the proposed schemes for different values of $\Delta$. We can see that all the schemes enjoy higher UDTs as $\Delta$ increases. This is because a higher $\Delta$ would enable the packets to be generated at more scattered time slots, which indeed leads to less urgent scenarios. Further, we can observe that the performance advantage of the proposed schemes over the optimal static scheme becomes less noticeable when $\Delta$ increases. This phenomenon can be attributed to the fact that the proposed schemes are designed based on the dynamic optimality under frame-synchronized traffic, thus being less robust to the introduction of time offsets.

Obviously, our POMDP formulation can be modified to support dynamic-varying $N$ if its probability distribution is known, which makes Algorithm 1 still applicable. Assuming $N$ is uniformly and randomly chosen from all integers in an interval with the mean 20, Fig. 6(d) indicates that $\boldsymbol{\pi}^{\text{simQ}}$ outperforms other policies for different interval lengths: 3.73% – 4.61% improvement over $\boldsymbol{\pi}^{\text{myo}}$ and 9.23% – 10.62% im-

[10]This method is simple but does not utilize the information from the transmission outcomes to accelerate the estimate process. Then, by viewing the problem with unknown $\lambda$ as a POMDP formulation with an unknown parameter, we can also use reinforcement learning algorithms to learn the value of $\lambda$ by directly interacting with the environment, which, however, are beyond the scope of this paper and will be considered as a direction of our future study.

[11]To handle with such unsynchronized traffic, a standard way is to develop an optimal scheme based on the theory of decentralized MDP. However, solving a decentralized MDP is in general NEXP-complete [41]. Hence, an appropriate practical scheme needs to be further designed, which is our ongoing work.

provement over $\boldsymbol{\pi}^{\text{sta}}$ under the collision channel, and $3.75\%$ – $4.39\%$ improvement over $\boldsymbol{\pi}^{\text{myo}}$ and $6.67\%$ – $8.12\%$ over $\boldsymbol{\pi}^{\text{sta}}$ under the SINR-based capture channel. As the interval length increases, we can see that all the schemes suffer lower UDTs. This is because the contention scenario becomes more unpredictable and varied.

## VIII. Conclusions

We investigated the random access design in uplink IoT systems for urgency-constrained frame-synchronized traffic. Built on the theories of MDP and POMDP, we generalized prior studies on this issue to seek optimal policies by considering a general ALOHA-like protocol, a general urgency function, and a general SPR channel. For the idealized environment, we proved the optimality of the myopic policy. For the realistic environment, we showed the myopic policy is in general not optimal, and proposed two practical policies that utilize the special properties of our problem. Simulation results showed that the proposed schemes outperform the state-of-the-art schemes under a wide range of network configurations. Our modeling approach can be easily extended to consider a general MPR channel (via modifying the state transition function, observation function, and reward function), a general frame-synchronized traffic model (via modifying the initial belief), and multi-slot packets (via incorporating the remaining time of current transmission into the state).

Instantaneous feedback is a key assumption in our work, which is reasonable when ACK/NACK transmission time is negligible compared with the packet transmission time. In our future work, we will relax this assumption to consider deferred-feedback cases, which would lead to more complicated modeling than that for the instantaneous-feedback case.

## Appendix A
## Proof of Inequality (14)

First, we shall prove $U_t^*(1)=\sigma_1\Gamma_D$ by induction on $t$ from $t=D$ down to 1. When $t=D$, by Eq. (10), we have $U_D^*(1)=\sigma_1\Gamma_D$. When $t\in\mathcal{T}\setminus\{D\}$, we assume $U_{t+1}^*(1)=\sigma_1\Gamma_D$. By Eqs. (1), (2), and (11), we have

$$\begin{aligned}U_t^{\diamond}(1,p)&=r_t(1,p)+\sum_{n'\in\mathcal{N}}\beta_t(n',1,p)U_{t+1}^*(n')\\&=\Gamma_t\sigma_1 p+(1-\sigma_1 p)U_{t+1}^*(1)\\&=(\Gamma_t-\sigma_1\Gamma_D)\sigma_1 p+\sigma_1\Gamma_D.\end{aligned}\tag{31}$$

Taking the derivative of Eq. (31) with respect to $p$ derives that $\frac{\mathrm{d}}{\mathrm{d}p}U_t^{\diamond}(1,p)=(\Gamma_t-\sigma_1\Gamma_D)\sigma_1$. As $\Gamma_t<\sigma_1\Gamma_{t+1}$, we obtain $\Gamma_t<\sigma_1^{D-t}\Gamma_D\le\sigma_1\Gamma_D$ for each $t\in\mathcal{T}\setminus\{D\}$ and, thus, $U_t^{\diamond}(1,p)$ attains its maximum when $p=0$, i.e., $U_t^*(1)=\sigma_1\Gamma_D$. So, we have $U_t^*(1)=\sigma_1\Gamma_D$ for each $t\in\mathcal{T}$.

Next, we shall prove $U_t^{\text{myo}}(1)<\sigma_1\Gamma_D$ for each $t\in\mathcal{T}\setminus\{D\}$. For each $t\in\mathcal{T}$, by Eqs. (2) and (7), we have $\widehat{\pi}_t^{\text{myo}}(1)=1$. When $t=D$, by Eq. (8), we have $U_D^{\text{myo}}(1)=\sigma_1\Gamma_D$. When $t\in\mathcal{T}\setminus\{D\}$, by Eqs. (1), (2), and (8), we have

$$\begin{aligned}U_t^{\text{myo}}(1)&=r_t\big(1,\widehat{\pi}_t^{\text{myo}}(1)\big)+\sum_{n'\in\mathcal{N}}\beta_t\big(n',1,\widehat{\pi}_t^{\text{myo}}(1)\big)U_{t+1}^{\text{myo}}(n')\\&=\Gamma_t\sigma_1+(1-\sigma_1)U_{t+1}^{\text{myo}}(1).\end{aligned}\tag{32}$$

Further, by $\Gamma_t<\sigma_1^{D-t}\Gamma_D$, recursively using Eq. (32) yields

$$\begin{aligned}U_t^{\text{myo}}(1)&=\sigma_1\Gamma_t+\sigma_1(1-\sigma_1)\Gamma_{t+1}+\cdots\\&\quad+\sigma_1(1-\sigma_1)^{D-t-1}\Gamma_{D-1}+(1-\sigma_1)^{D-t}U_D^{\text{myo}}(1)\\&=\sigma_1\Gamma_t+\sigma_1(1-\sigma_1)\Gamma_{t+1}+\cdots\\&\quad+\sigma_1(1-\sigma_1)^{D-t-1}\Gamma_{D-1}+\sigma_1(1-\sigma_1)^{D-t}\Gamma_D\\&<\sigma_1^{D-t+1}\Gamma_D+\sigma_1^{D-t}(1-\sigma_1)\Gamma_D+\cdots\\&\quad+\sigma_1^2(1-\sigma_1)^{D-t-1}\Gamma_D+\sigma_1(1-\sigma_1)^{D-t}\Gamma_D\\&=\sigma_1\big(\sigma_1^{D-t}+\sigma_1^{D-t-1}(1-\sigma_1)+\cdots\\&\quad+\sigma_1(1-\sigma_1)^{D-t-1}+(1-\sigma_1)^{D-t}\big)\Gamma_D\\&\le\sigma_1\big(\sigma_1+(1-\sigma_1)\big)^{D-t}\Gamma_D=\sigma_1\Gamma_D.\end{aligned}$$

## Appendix B
## Proof of Inequality (15)

In this proof, the MDP definitions and notation are given based on the threshold-based MPR channel assumption. So, we have

$$\begin{aligned}&\beta_t(n',n,p)\\&=\begin{cases}\binom{n}{n-n'}p^{n-n'}(1-p)^{n'}, & \text{if } 0<n-n'\le\gamma,\\ 1-\sum_{k=1}^{\min(n,\gamma)}\binom{n}{k}p^k(1-p)^{n-k}, & \text{if } n-n'=0,\\ 0, & \text{otherwise,}\end{cases}\end{aligned}\tag{33}$$

for each $t\in\mathcal{T}\setminus\{D\}$, each $n,n'\in\mathcal{N}$, and each $p\in[0,1]$, and

$$r_t(n,p)=\sum_{k=1}^{\min(n,\gamma)}k\binom{n}{k}p^k(1-p)^{n-k},\tag{34}$$

for each $t\in\mathcal{T}$, each $n\in\mathcal{N}$, and each $p\in[0,1]$. We also have

$$U_t^*(n)=U_t^{\text{myo}}(n)=n,\ \forall t\in\mathcal{T},\ \forall n\in\{0,1,\ldots,\gamma\},\tag{35}$$

$$U_D^*(n)=U_D^{\text{myo}}(n),\ \forall n\in\mathcal{N}.\tag{36}$$

Then, given arbitrary $D\ge1$ and $2\le\gamma<N$, for each $t\in\mathcal{T}$, by Eq. (34), we have

$$\begin{aligned}r_t(\gamma+1,p)&=(\gamma+1)\sum_{k'=0}^{\gamma-1}\binom{\gamma}{k'}p^{k'+1}(1-p)^{\gamma-k'}\\&=(\gamma+1)(p-p^{\gamma+1}).\end{aligned}$$

Since $r_t(\gamma+1,p)>r_t(\gamma+1,0)=r_t(\gamma+1,1)=0$ if $p\in(0,1)$, we know that the continuous function $r_t(\gamma+1,p)$ has a local maximum at $\widehat{\pi}_t^{\text{myo}}(\gamma+1)$ lying in $(0,1)$. As $\frac{\mathrm{d}}{\mathrm{d}p}r_t(\gamma+1,p)=(\gamma+1)\big(1-(\gamma+1)p^{\gamma}\big)$ always exists at $p\in(0,1)$, by the Fermat's Theorem, $\widehat{\pi}_t^{\text{myo}}(\gamma+1)$ is a solution of $\frac{\mathrm{d}}{\mathrm{d}p}r_t(\gamma+1,p)=0$, i.e., $\widehat{\pi}_t^{\text{myo}}(\gamma+1)=(\gamma+1)^{-\frac{1}{\gamma}}$.

We further investigate the monotonicity of the sequence $\{\widehat{\pi}_t^{\text{myo}}(\gamma+1)\}_{\gamma=2}^{N-1}$. Let $f(x)\triangleq(x+1)^{-\frac{1}{x}}$, $x\in[2,+\infty)$ such that $f(\gamma)=\widehat{\pi}_t^{\text{myo}}(\gamma+1)$ for all $2\le\gamma<N$. Taking the derivative of $f(x)$ with respect to $x$ derives that $\frac{\mathrm{d}}{\mathrm{d}x}f(x)=x^{-2}(x+1)^{-\frac{x+1}{x}}\big((x+1)\ln(x+1)-x\big)$. Let $g(x)\triangleq(x+1)\ln(x+1)-x$, $x\in[2,+\infty)$. Since $\frac{\mathrm{d}}{\mathrm{d}x}g(x)=\ln(x+1)>0$, the function $g(x)$ is strictly increasing on $[2,+\infty)$. Thus, we obtain that $g(x)\ge g(2)=$

$3\ln 3 - 2 > 0$ and $\frac{\mathrm{d}}{\mathrm{d}x} f(x) = x^{-2}(x+1)^{-\frac{x+1}{x}} g(x) > 0$. As the function $f(x)$ is strictly increasing on $[2, +\infty)$, the sequence $\{\widehat{\pi}_t^{\text{myo}}(\gamma+1)\}_{\gamma=2}^{N-1}$ is strictly increasing. So, for all $2 \le \gamma < N$, we have

$$\widehat{\pi}_t^{\text{myo}}(\gamma+1) \ge \widehat{\pi}_t^{\text{myo}}(3) > \frac{1}{2}. \tag{37}$$

Next, given arbitrary $D \ge 2$ and $2 \le \gamma < N$, when $t \in \mathcal{T} \setminus \{D\}$, by Eqs. (11), (33), (34), and (35), we have

$$\begin{aligned}
&U_t^{\diamond}(\gamma+1, p) = r_t(\gamma+1, p) + \sum_{n' \in \mathcal{N}} \beta_t(n', \gamma+1, p) U_{t+1}^*(n') \\
&= \sum_{k=1}^{\gamma} k \binom{\gamma+1}{k} p^k (1-p)^{\gamma+1-k} \\
&\quad + \sum_{k=1}^{\gamma} \binom{\gamma+1}{k} p^k (1-p)^{\gamma+1-k} U_{t+1}^*(\gamma+1-k) \\
&\quad + \left(1 - \sum_{k=1}^{\gamma} \binom{\gamma+1}{k} p^k (1-p)^{\gamma+1-k}\right) U_{t+1}^*(\gamma+1) \\
&= (\gamma+1)\big(1 - p^{\gamma+1} - (1-p)^{\gamma+1}\big) \\
&\quad + \big(p^{\gamma+1} + (1-p)^{\gamma+1}\big) U_{t+1}^*(\gamma+1) \\
&= \big(\gamma + 1 - U_{t+1}^*(\gamma+1)\big) \\
&\quad \times \big(1 - p^{\gamma+1} - (1-p)^{\gamma+1}\big) + U_{t+1}^*(\gamma+1).
\end{aligned} \tag{38}$$

Taking the derivative of $U_t^{\diamond}(\gamma+1, p)$ with $p$ derives that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}p} U_t^{\diamond}(\gamma+1, p) \\
&= \big(\gamma + 1 - U_{t+1}^*(\gamma+1)\big)\big(1 - p^{\gamma+1} - (1-p)^{\gamma+1}\big)' \\
&= (\gamma+1)\big(\gamma + 1 - U_{t+1}^*(\gamma+1)\big)\big((1-p)^{\gamma} - p^{\gamma}\big).
\end{aligned}$$

Let $f(p) \triangleq (1-p)^{\gamma} - p^{\gamma}$, $p \in [0,1]$. The derivative of $f(p)$ with respect to $p$ is $\frac{\mathrm{d}}{\mathrm{d}p} f(p) = -\gamma((1-p)^{\gamma-1} + p^{\gamma-1}) < 0$ for $p \in [0,1]$. As $f(p)$ is strictly decreasing on $[0,1]$, $p = \frac{1}{2}$ is the only solution of $f(p) = 0$. By the MDP definitions, we know $\gamma + 1 - U_t^*(\gamma+1) > 0$ for each $t \in \mathcal{T}$. Since $\frac{\mathrm{d}}{\mathrm{d}p} U_t^{\diamond}(\gamma+1, p) = (\gamma+1)\big(\gamma + 1 - U_{t+1}^*(\gamma+1)\big) f(p)$, $U_t^{\diamond}(\gamma+1, p)$ is strictly increasing on $p \in [0, \frac{1}{2})$ and strictly decreasing on $p \in (\frac{1}{2}, 1]$. So, we have $\widehat{\pi}_t^*(\gamma+1) = \frac{1}{2}$ for each $t \in \mathcal{T} \setminus \{D\}$.

For each $t \in \mathcal{T} \setminus \{D\}$, by Inequality (37), we have $\widehat{\pi}_t^*(\gamma+1) < \widehat{\pi}_t^{\text{myo}}(\gamma+1)$ and $U_t^{\text{myo}}(\gamma+1) < U_t^*(\gamma+1)$. Hence, we complete the proof for Inequality (15).

## Appendix C
## Proof of Inequality (23)

Let $f(\mathbf{b}, p) \triangleq \sum_{n \in \mathcal{N}} b(n) r_t(n, p)$. Then $f(\mathbf{b}, p) = (\sigma_2 - 2) b(2) p^2 + (b(1) + 2b(2)) p$ and $\frac{\mathrm{d}}{\mathrm{d}p} f(\mathbf{b}, p) = (2\sigma_2 - 4) b(2) p + b(1) + 2b(2)$. By the assumption $0 < b(2) \le \frac{1}{2-2\sigma_2} b(1)$, $f(\mathbf{b}, p)$ attains its maximum only when $p = 1$, indicating $\pi_{D-1}^{\text{myo}}(\mathbf{b}) = 1$. By Eq. (20), we have

$$\begin{aligned}
V_{D-1}^{\text{myo}}(\mathbf{b}) &= b(1) + \sigma_2 b(2) \\
&\quad + \chi_{D-1}(0, \mathbf{b}, 1) V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, 1, 0)\big) \\
&\quad + \chi_{D-1}(1, \mathbf{b}, 1) V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, 1, 1)\big) \\
&\quad + \chi_{D-1}(2, \mathbf{b}, 1) V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, 1, 2)\big),
\end{aligned}$$

where

$$\begin{aligned}
&\chi_{D-1}(0, \mathbf{b}, 1) = b(0), \\
&\theta_{D-1}(\mathbf{b}, 1, 0) = [1, 0, 0], \\
&\chi_{D-1}(1, \mathbf{b}, 1) = b(1) + \sigma_2 b(2), \\
&\theta_{D-1}(\mathbf{b}, 1, 1) = \frac{1}{\chi_{D-1}(1, \mathbf{b}, 1)} [b(1), \sigma_2 b(2), 0], \\
&\chi_{D-1}(2, \mathbf{b}, 1) = (1 - \sigma_2) b(2), \\
&\theta_{D-1}(\mathbf{b}, 1, 2) = [0, 0, 1], \\
&V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, 1, 0)\big) = 0, \\
&V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, 1, 1)\big) = \frac{\sigma_2 b(2)}{\chi_{D-1}(1, \mathbf{b}, 1)}, \\
&V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, 1, 2)\big) = \frac{1}{2 - \sigma_2}.
\end{aligned}$$

So, we have $V_{D-1}^{\text{myo}}(\mathbf{b}) = b(1) + \frac{-2\sigma_2^2 + 3\sigma_2 + 1}{2 - \sigma_2} b(2)$.

For the ease of notation, let $\rho = \frac{1}{2-\sigma_2}$. By Eqs. (21) and (22), we have

$$\begin{aligned}
V_{D-1}^{\diamond}(\mathbf{b}, \rho) &= b(1)\rho + b(2)\big(2\rho(1-\rho) + \sigma_2 \rho^2\big) \\
&\quad + \chi_{D-1}(0, \mathbf{b}, \rho) V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, \rho, 0)\big) \\
&\quad + \chi_{D-1}(1, \mathbf{b}, \rho) V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, \rho, 1)\big) \\
&\quad + \chi_{D-1}(2, \mathbf{b}, \rho) V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, \rho, 2)\big),
\end{aligned}$$

where

$$\begin{aligned}
&\chi_{D-1}(0, \mathbf{b}, \rho) = b(0) + b(1)(1-\rho) + b(2)(1-\rho)^2, \\
&\theta_{D-1}(\mathbf{b}, \rho, 0) = \frac{1}{\chi_{D-1}(0, \mathbf{b}, \rho)} \\
&\qquad \times \big[b(0), b(1)(1-\rho), b(2)(1-\rho)^2\big], \\
&\chi_{D-1}(1, \mathbf{b}, \rho) = b(1)\rho + b(2)\big(2\rho(1-\rho) + \sigma_2\rho^2\big), \\
&\theta_{D-1}(\mathbf{b}, \rho, 1) = \frac{1}{\chi_{D-1}(1, \mathbf{b}, \rho)} \\
&\qquad \times \big[b(1)\rho, b(2)\big(2\rho(1-\rho) + \sigma_2\rho^2\big), 0\big], \\
&\chi_{D-1}(2, \mathbf{b}, \rho) = b(2)(1-\sigma_2)\rho^2, \\
&\theta_{D-1}(\mathbf{b}, \rho, 2) = [0, 0, 1], \\
&V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, \rho, 0)\big) = \frac{b(1)(1-\rho) + \sigma_2 b(2)(1-\rho)^2}{\chi_{D-1}(0, \mathbf{b}, \rho)}, \\
&V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, \rho, 1)\big) = \frac{2\rho(1-\rho) + \sigma_2\rho^2}{\chi_{D-1}(1, \mathbf{b}, \rho)}, \\
&V_D^{\text{myo}}\big(\theta_{D-1}(\mathbf{b}, \rho, 2)\big) = \rho.
\end{aligned}$$

So, we have

$$V_{D-1}^{\diamond}\big(\mathbf{b}, \frac{1}{2-\sigma_2}\big) = b(1) + \frac{-\sigma_2^4 + 4\sigma_2^3 - 3\sigma_2^2 - 7\sigma_2 + 9}{(2-\sigma_2)^3} b(2).$$

Comparing $V_{D-1}^{\text{myo}}(\mathbf{b})$ and $V_{D-1}^{\diamond}\big(\mathbf{b}, \frac{1}{2-\sigma_2}\big)$, we have

$$\begin{aligned}
&V_{D-1}^{\diamond}\big(\mathbf{b}, \frac{1}{2-\sigma_2}\big) - V_{D-1}^{\text{myo}}(\mathbf{b}) \\
&= \frac{(1-\sigma_2)^2(\sigma_2^2 - 5\sigma_2 + 5)}{(2-\sigma_2)^3} b(2) > 0,
\end{aligned}$$

for each $\sigma_2 \in [0, 1)$.

## REFERENCES


[1] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[2] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, and B. Vucetic, "High-reliability and low-latency wireless communication for Internet of Things: Challenges, fundamentals, and enabling technologies," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7946–7970, 2019.

[3] M. Luvisotto, Z. Pang, and D. Dzung, "High-performance wireless networks for industrial control applications: New targets and feasibility," *Proc. IEEE*, vol. 107, no. 6, pp. 1074–1093, 2019.

[4] Y. Gao and L. Dai, "Random access: Packet-based or connection-based?" *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2664–2678, 2019.

[5] *Study on New Radio (NR) Access Technology*. document TS 38.912 v16.0.0, 3GPP, Jul. 2020.

[6] Y. H. Bae, "Analysis of optimal random access for broadcasting with deadline in cognitive radio networks," *IEEE Commun. Lett.*, vol. 17, no. 3, pp. 573–575, 2013.

[7] N. Nomikos, N. Pappas, T. Charalambous, and Y.-A. Pignolet, "Deadline-constrained bursty traffic in random access wireless networks," in *Proc. IEEE 19th SPAWC*, 2018, pp. 1–5.

[8] Y. Zhang, Y.-H. Lo, F. Shu, and J. Li, "Achieving maximum reliability in deadline-constrained random access with multiple-packet reception," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 5997–6008, 2019.

[9] V. Naware, G. Mergen, and L. Tong, "Stability and delay of finite-user slotted aloha with multipacket reception," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2636–2656, 2005.

[10] L. Deng, J. Deng, P.-N. Chen, and Y. S. Han, "On the asymptotic performance of delay-constrained slotted ALOHA," in *Proc. IEEE ICCCN*, 2018, pp. 1–8.

[11] H. ElSawy, "Characterizing IoT networks with asynchronous time-sensitive periodic traffic," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1696–1700, 2020.

[12] Y. H. Bae, "Queueing analysis of deadline-constrained broadcasting in wireless networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1782–1785, 2015.

[13] Y. H. Bae, "Modeling timely-delivery ratio of slotted Aloha with energy harvesting," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1823–1826, 2017.

[14] E. Fountoulakis, T. Charalambous, N. Nomikos, A. Ephremides, and N. Pappas, "Information freshness and packet drop rate interplay in a two-user multi-access channel," *J. Commun. Netw.*, vol. 24, no. 3, pp. 357–364, 2022.

[15] Y. H. Bae and J. W. Baek, "Age of information and throughput in random access-based IoT systems with periodic updating," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 821–825, 2022.

[16] L. Zhao, X. Chi, L. Qian, and W. Chen, "Analysis on latency-bounded reliability for adaptive grant-free access with multipackets reception (MPR) in URLLCs," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 892–895, 2019.

[17] R. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 3, pp. 323–328, 1987.

[18] H. Wu, C. Zhu, R. J. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA using access history for event-driven M2M communications," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1904–1917, 2013.

[19] W. T. Toor, J.-B. Seo, and H. Jin, "Online control of random access with splitting," in *Proc. ACM Mobihoc*, 2020, pp. 61–70.

[20] S.-W. Jeon and H. Jin, "Online estimation and adaptation for random access with successive interference cancellation," *IEEE Trans. Mobile Comput.*, 2022, doi: 10.1109/TMC.2022.3179240.

[21] W. T. Toor, J.-B. Seo, and H. Jin, "Practical splitting algorithm for multi-channel slotted random access systems," *IEEE Trans. Mobile Comput.*, vol. 19, no. 12, pp. 2863–2873, 2020.

[22] O. T. Yavascan and E. Uysal, "Analysis of slotted ALOHA with an age threshold," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1456–1470, 2021.

[23] X. Chen, K. Gatsis, H. Hassani, and S. S. Bidokhti, "Age of information in random access channels," *IEEE Trans. Inf. Theory*, vol. 68, no. 10, pp. 6548–6568, 2022.

[24] J. Sun, Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Closed-form Whittle's index-enabled random access for timely status update," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1538–1551, 2019.

[25] H. H. Yang, A. Arafa, T. Q. Quek, and H. V. Poor, "Spatiotemporal analysis for age of information in random access networks under last-come first-serve with replacement protocol," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2813–2829, 2021.

[26] G. del Angel and T. L. Fine, "Optimal power and retransmission control policies for random access systems," *IEEE/ACM Trans. Netw.*, vol. 12, no. 6, pp. 1156–1166, 2004.

[27] K. Cohen and A. Leshem, "Distributed game-theoretic optimization and management of multichannel ALOHA networks," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1718–1731, 2015.

[28] A. Biason, S. Dey, and M. Zorzi, "A decentralized optimization framework for energy harvesting devices," *IEEE Trans. Mobile Comput.*, vol. 17, no. 11, pp. 2483–2496, 2018.

[29] A. Fu and M. Mazo, "Traffic models of periodic event-triggered control systems," *IEEE Trans. Autom. Control*, vol. 64, no. 8, pp. 3453–3460, 2018.

[30] *A 5G traffic model for industrial use cases*. White Paper, 5G Alliance for Connected Industries and Automation, 2019.

[31] Y. Zhang, A. Gong, Y.-H. Lo, J. Li, F. Shu, and W. S. Wong, "Generalized $p$-persistent CSMA for asynchronous multiple-packet reception," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6966–6979, 2019.

[32] K. Wang, L. Chen, and J. Yu, "On optimality of myopic policy in multi-channel opportunistic access," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 677–690, 2016.

[33] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers *et al.*, "Practical lessons from predicting clicks on ads at Facebook," in *Proc. 8th Int. Workshop Data Min. Online Adv.*, 2014, pp. 1–9.

[34] A. Silik, M. Noori, W. A. Altabey, J. Dang, R. Ghiasi, and Z. Wu, "Optimum wavelet selection for nonparametric analysis toward structural health monitoring for processing big data from sensor network: A comparative study," *Struct. Health Monit.*, vol. 21, no. 3, pp. 803–825, 2022.

[35] Y. Wang, M. C. Vuran, and S. Goddard, "Analysis of event detection delay in wireless sensor networks," in *Proc. IEEE INFOCOM*, 2011, pp. 1296–1304.

[36] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[37] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon," *Oper. Res.*, vol. 21, no. 5, pp. 1071–1088, 1973.

[38] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*. SIAM, 2015.

[39] M. Hauskrecht, "Value-function approximations for partially observable Markov decision processes," *J. Artif. Intell. Res.*, vol. 13, pp. 33–94, 2000.

[40] Č. Stefanović, M. Momoda, and P. Popovski, "Exploiting capture effect in frameless ALOHA for massive wireless random access," in *Proc. IEEE WCNC*, 2014, pp. 1762–1767.

[41] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of Markov decision processes," *Math. Oper. Res.*, vol. 27, no. 4, pp. 819–840, 2002.