2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)

Scalable Multi-Agent Learning for Situationally-Aware Multiple-Access and Grant-Free Transmissions

Zhiyuan Jiang*[†], Andrei Marinescu[‡], Luiz A. DaSilva[‡], *Fellow, IEEE*, Sheng Zhou[†], Zhisheng Niu[†], *Fellow, IEEE*
∗: Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China.
‡: CONNECT, Trinity College Dublin, Dublin 2, Ireland.
†: Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China.
Emails: zhiyjiang@foxmail.com, {marinesa, dasilval}@tcd.ie, {sheng.zhou, niuzhs}@tsinghua.edu.cn

***Abstract*—In this paper, we present Situationally-aware Multiple-Access and gRant-free Transmissions (SMART) to address the low-latency multiple access issue in wireless uplinks with massive machine-type communications (mMTC). SMART is smart in a distributed manner, as terminals are trained to be situationally aware. The solution is based on a distributed reinforcement learning framework which is capable of dealing with diversified quality-of-service requirements. As such, terminals can make informed transmission decisions by themselves, e.g., by taking into account the urgency of their packets and system load. In this way, the effective number of concurrent access terminals is significantly reduced while maintaining the system performance. Compared with conventional contention-based random access schemes, SMART has significant advantages. Building upon our previous work [1], this work presents the first SMART algorithm that is scalable to massive terminals (hundreds) and stable with near-optimal performance.**

I. Introduction

5G and beyond wireless networks are promising to achieve near-instantaneous communications (1-10 ms) for high-density (10^6 devices/km^2) machine-type terminals to enable novel applications such as autonomous driving, Industrial Internet of Things (IIoT) and intelligent health care. Based on the current standardization progress and research trend, this represents a feasible target in the wireless downlinks, wherein ultra-Reliable Low-Latency Communications (uRLLC) traffic can be preemptively overlapped with enhanced Mobile Broadband (eMBB) traffic in the time-frequency domain [2]; that is, the uRLLC traffic can be scheduled immediately upon arrival on top of the existing eMBB traffic, which significantly reduces the scheduling delay. The root reason of the feasibility of this approach is the *centralized scheduling* nature in wireless downlinks, where the scheduling decisions can be made centrally and the corresponding control signaling is transmitted to the terminals occupying the same physical resources, e.g., by the Physical Downlink Control Channel (PDCCH) in 4G/5G.

However, the wireless medium access delay in the uplinks is much more challenging due to the *decentralized* nature of terminals and their stochastic (sometimes bursty) traffic demands, e.g., for safety massages in autonomous driving. In grant-based uplink access schemes, which are—as of today with 3GPP Rel 15 [3]—still the selected access schemes in 5G, the access delay is dominated by the time of waiting for an uplink access opportunity, i.e., Scheduling Request (SR). Typically, the uplink access delay is around 20-30 ms, which is insufficient for uRLLC traffic. In addition, this excludes the extra access delay when SR is blocked due to bursty traffic demand. Towards this end, *grant-free* [4] access schemes are proposed to enable arrive-and-go transmissions in the wireless uplinks. Specifically, terminals with packets to transmit undergo a contention process, and winners directly transmit their packets without having to wait for downlink scheduling grants. Grant-free transmissions significantly reduce the uplink access delay when traffic load is low [4].

On the other hand, inherited from its contention-based mechanism, the grant-free transmission scheme suffers severe performance degradation when the number of concurrent terminals is large. To address this issue, extensive efforts have been made, which can be categorized by either aiming to prioritize the access [5], [6], or to enhance the physical access capabilities [4], [7]. For the purpose of access prioritization, the Extended Access Barring (EAB) technique has been suggested [6]. This technique randomly selects a certain set of terminals to transmit by broadcasting a threshold by the access point and each terminal generating a random number. Prioritized Random Access (PRA) prescribes different access backoff mechanisms and resource allocation for terminals with pre-defined priority classes, e.g., human-to-human traffic, low- and high-priority traffic. Sparse-Code Multiple-Access (SCMA) [4] and beamforming [7] techniques can be regarded as aiming to enhance the number of concurrent terminals by strengthening the receiver capability in the code and spatial domains, respectively.

In this paper, we propose a novel multi-agent learning-based uplink access framework for real-time signal transmissions with diversified Quality-of-Service (QoS) requirements, namely Situationally-aware Multiple-Access and gRant-free Transmissions (SMART). SMART is smart distributedly, in the sense that terminals' access probabilities, or equivalently backoff window sizes, are based on their individual situa-

978-1-5386-6528-2/19/$31.00 ©2019 IEEE

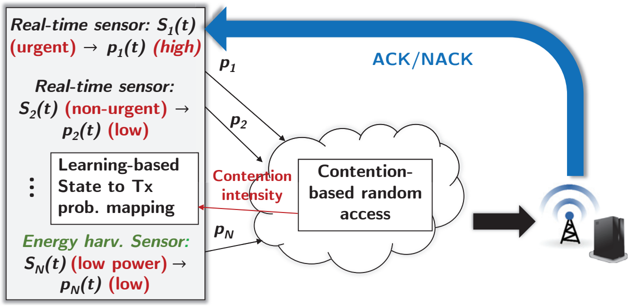

Fig. 1. SMART framework, wherein heterogeneous QoS requirements can be simultaneously satisfied by learning-based access strategies according to terminals' states.

tions (formally defined as their Markov states) and learned through a multi-agent reinforcement learning framework. Such an approach is promising in real-world applications wherein terminals collect sensory information about the physical world; the urgency of the information is time-varying and known to the terminals such that terminals can decide the transmission strategy dynamically thereby. Based on SMART, the *effective* number of concurrent terminals can be reduced, hence the uplink access delay. In addition, heterogeneous QoS can be incorporated in this framework since no restrictions on the Markov states are enforced.

Related work stems from the recent emerging research interests of various latency, or age, metrics, e.g., Age of Information (AoI) [8], Age of Synchronization (AoS) [9], Age upon Decision (AuD) [10] and Inter-Delivery Time (IDT) [11]. These age metrics are proposed for different application scenarios, whereas sharing a common feature that a terminal which is aiming to optimize its corresponding metric has dynamic, time-varying transmission urgency. In our previous works [12]–[15], we have derived closed-form Whittle's index based policies and decentralized scheduling schemes exclusively for AoI optimizations. Based on our knowledge, this is the first work that achieves scalable Multi-Agent Reinforcement Learning (MARL)-based random access for heterogeneous QoS requirements.

II. System Model and Problem Formulation

The considered system model is shown in Fig. 1. A multiaccess network is considered wherein an information fusion center (FC) collects information from N distributed terminals with possibly heterogeneous QoS requirements. Time is slotted and we assume the terminals are synchronized, which can be realized by assuming that the terminals have maintained synchronization by receiving the Primary Synchronization Signal (PSS) from the FC but with no scheduling grants. The transmission model in the uplink is collision-based. A transmission frame consists of a data slot and several contention mini-slots. A data slot (with length T_{s}) is prefixed by several contention mini-slots (with length δ). In 5G New Radio (NR), the concept of mini-slots [7] is introduced which is the minimum scheduling time unit, occupying as short as one OFDM symbol. In addition to the scalable numerology of NR, wherein one slot, consisting of 14 OFDM symbols, can be 0.125 ms with 120 KHz subcarrier spacing (SCS), each mini-slot can be quite short ($\delta = 1/56$ ms or lower with larger SCS). We assume a p-persistent Carrier-Sense Multiple Access (CSMA) framework [16], whereby terminal-i transmits with a probability p_i in each contention mini-slot when it senses the channel is idle; otherwise it stays silent. Note that, different from homogeneous p-persistent CSMA, the persistent levels of terminals can be different, i.e., p_i differs among terminals—in this way, the terminals can be situationally-aware and thereby choosing appropriate p_i. Note that in the Q-CSMA scheme [17], p_i is determined by the queue length of each terminal, which however only applies to throughput optimizations. Based on the definition of p-persistent CSMA, the terminal who has won the contention transmits in the following data slot and the others sense that the channel is busy and stay silent. After a data slot, the FC feeds back an acknowledgment packet (ACK) indicating successful reception; otherwise a NACK packet is fed back. Note that a p-persistent CSMA protocol closely approximates the IEEE 802.11 CSMA protocol which employs uniform backoff counters and binary exponential back, if p and the backoff window size are chosen such that the average backoff intervals of the two protocols are identical.

In this work, we only consider the access to a single channel. By using e.g., SCMA, beamforming and multiple frequency sub-channels, it is certainly possible to enable Multiple Packets Reception (MPR) simultaneously. A straightforward approach to extend from single-channel to MPR is to let the terminals choose one channel randomly or uniformly pre-allocate the channels.

Problem Formulation: The objective is to maximize the overall utility of all terminals over time, i.e.,

$$\max_{p_i(t), i=1,\cdots,N} \sum_{i=1}^{N} \omega_i \overline{U}_i, \;\text{ where } p_i(t) \text{ only depends on } \boldsymbol{s}_i(t), \tag{1}$$

wherein $\overline{U}_i \triangleq \liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} U_i(t)$, the utility function of terminal-i at time t is denoted by $U_i(t)$ which reflects the QoS requirements and will be instantiated in the following section. The terminal weight is denoted by ω_i. In each time slot, the terminals choose their transmission probability $p_i(t)$, depending only on their own situation, i.e., Markov states $\boldsymbol{s}_i(t)$, by learning-based approaches that will be specified later—this decentralized approach is especially important in future massive IoT systems, in order to avoid the prohibitive high signaling overhead for centralized scheduling methods.

General Situation (Markov State) Characterizations: The situation that terminal-i is in at time t is denoted by its Markov state, i.e., $\boldsymbol{s}_i(t)$, which is a real-valued vector. The definition is quite general, and encompasses arbitrary state space and transition dynamics. Note that even for non-Markov states, we can define the state as the state concatenation of several historical states as approximated Markov states.

III. Proposed Scalable MARL Approach

When considering the general Markov state definition in the previous section, it is quite challenging to find a universal analytical solution for various types of states and QoS requirements. For example, our previous works [13]–[15] found that, even for AoI optimizations alone, this can be quite challenging, let alone for the co-existence of diversified types of states. In view of this, we resort to the MARL framework. RL is a model-free control mechanism and therefore is applicable to arbitrary types of states and state transitions. The unique property of this problem is that each terminal, or agent in RL terms, can only observe its own states and make decisions thereby—this is referred to as a *partially observable identical payoff stochastic game* (POIPSG) [18], which is used to model a problem wherein multiple agents learn simultaneously with a single objective (total utility functions) and observations of only local state information. The game notation reflects the interplay among terminals which is quite different from the conventional static environment setting in RL as agents can interfere with each other while learning, and is thus named MARL.

The MARL nature of the problem poses significant challenges to find a scalable and stable solution. The challenging part, which is well-known for MARL problems, is that each agent only observes its own states evolution over time, and its environment involves the actions of other agents, thus making it non-stationary—like trying to learn from a moving target. At the same time, all agents try to learn a good policy, and it is very hard to design a stable and scalable (massive number of terminals are common in IoT) solution. As shown in the following sections, we have tried several approaches, including state-of-the-art MARL schemes, and find that the considered problem is not a trivial task. In the following, we describe these approaches and propose a novel scheme which combines the idea of Whittle's index and RL, achieving near-optimal performances consistently in various scenarios.

A. Transmission Tax Based Decoupled MARL Approach

First, we will introduce the proposed approach, namely Transmission Tax based decoupled MARL approach (TT). TT is based on the idea of decoupled RL training to avoid the convergence issue introduced by the interplay among agents in MARL. The challenging part is: How to ensure that the trained policy using decoupled RL training also works well under the multi-agent setting? In particular, if we naively train each agent separately with the objective of optimizing its own utility, then all agents would become selfish and putting them together, the channel would be jammed all the time because no agents are trained to cooperate with others. We resolve this issue by introducing a universal transmission tax for all terminals when trained separately. That is, when an agent is trained, a transmission tax (i.e., a cost m) is added whenever the agent choose to transmit; when it chooses to stay silent, no tax is added. By doing this, agents are trained to be less selfish, and more conservative in transmissions, i.e., only when an agent is in a situation where it has a high-value packet would it actually transmits since otherwise the transmission tax would surpass the value of a transmission. Surprisingly and unsurprisingly, this approach works well in various scenarios. The surprising part is the simple heuristics behind TT; the unsurprising part is that TT is in fact based on the idea of Whittle's index which is widely known to be a near-optimal approach for this kind of problems, specifically restless multi-armed bandit (RMAB).

The connection with Whittle's index approach is illustrated as follows. Based on the Whittle's index approach, the utility maximization scheduling problem is decomposed to N subproblems, where each subproblem can be formulated based on the Bellman optimality equations (average cost with infinite-horizon and relative cost-to-go functions [19]) as

$$f(\boldsymbol{s}) + \hat{J}^* = \min \left\{ \begin{array}{l} \mathcal{R}_{\boldsymbol{s}}^{(0)} + \sum_{\boldsymbol{s}'} \mathcal{P}_{\boldsymbol{s}\boldsymbol{s}'}^{(0)} f(\boldsymbol{s}'), \\ m + \mathcal{R}_{\boldsymbol{s}}^{(1)} + \sum_{\boldsymbol{s}'} \mathcal{P}_{\boldsymbol{s}\boldsymbol{s}'}^{(1)} f(\boldsymbol{s}') \end{array} \right\}, \quad (2)$$

wherein the top and bottom terms in the minimization operator represent the cost-to-go from state $\boldsymbol{s}$ onwards with the action of silent and transmit, respectively. The expected reward functions are denoted by $\mathcal{R}_{\boldsymbol{s}}^{(0)}$ and $\mathcal{R}_{\boldsymbol{s}}^{(1)}$ respectively for both actions; the transition matrices are denoted likewise. The relative cost-to-go function of state $\boldsymbol{s}$ and the average reward are denoted by $f(\boldsymbol{s})$ and $\hat{J}^*$ respectively. The terminal index is omitted for brevity while one should note that the reward functions, transition matrices and cost-to-go functions can all be different among terminals to reflect heterogeneous states and QoS, except for the transmission tax m which is identical among terminals.

There are two differences between TT and an exact Whittle's index policy. First, the scheduling decisions are centralized and deterministic in the Whittle's index policy, i.e., the index policy solves for the equivalent transmission tax for each state that makes the scheduling options of (2) equally good, and compares among terminals to find the one with the largest index. However, the decentralized transmission strategy considered in our formulation is stochastic, which is necessary in distributed settings. Secondly, the Whittle's index approach seeks for the maximum index (equivalent transmission tax) among terminals, while our approach lets all terminals share an identical m. In practice, one can argue that a scheme wherein a common transmission tax threshold is fixed, and terminals calculate their own decisions by solving the Markov Decision Process (MDP) of (2) and content for a transmission opportunity if the decision is to transmit (otherwise silent) is equivalent to the Whittle's index approach, if the transmission tax threshold is optimized.

Based on this intuition, we present TT in Algorithm 1. A golden search method (and hence the coefficients in Step 3), combined with Monte Carlo policy evaluations, is leveraged to find the optimal transmission tax; in Step 17 and 19, we can find the update of the transmission taxes accordingly. At each iteration, every terminal is trained separately based on the current transmission tax. We adopt the well-known deep Q-learning (DQN) [20] algorithm to accomplish the single-agent training tasks, and the model parameters for terminal-i

is denoted by $\boldsymbol{w}_i$. After single-agent training, all terminals use the current model parameters to participate in the multi-agent training phase, wherein each terminal would transmit with probability calculated based on [16] if it senses the channel is idle and its instantaneous DQN output is to transmit based on its current state; the FC feeds back an ACK/NACK after each data slot; the FC calculates the average reward (i.e., time-average utility) in each iteration and update the transmission tax accordingly. The process continues until the transmission tax converges. After the training, all terminals use the final model parameters for the random access procedure. In Step 3, the length of a contention mini-slot is denoted by δ, and ρ controls how many terminals contend in each iteration.

Algorithm 1: TT

```
1  **Initialization**:
2  Terminals: Initialize model parameters $\boldsymbol{w}_i$
     $(i = 1, \cdots, N)$ following the normal distribution.
3  FC: Use $m_{\mathsf{max}}$ and $m_{\mathsf{min}}$ to denote the maximum and
     minimum transmission taxes, respectively. Set
     $m_{\mathsf{min}} = 0$, $m_1 = \frac{3-\sqrt{5}}{2} m_{\mathsf{max}}$, $m_2 = \frac{\sqrt{5}-1}{2} m_{\mathsf{max}}$,
     $N_{\mathsf{target}} = \rho N$, and $p_{\mathsf{tx}} = \min\left\{\sqrt{\frac{2\delta}{T_{\mathsf{s}} N_{\mathsf{target}}^2}}, \frac{1}{N_{\mathsf{target}}}\right\}$.
4  for $k = \{1, 2\}$ do
5      $m = m_k$.
6      Decoupled Single-Agent Training:
7      for $i = 1 : N$ do
8          DQN training for terminal-$i$ to solve the MDP
             expressed in (2) with given $m$ to update their
             model parameters $\boldsymbol{w}_i$. The action output of
             DQNs is transmit or silent.
9      Multi-Agent Training:
10     for $t = 1 : T$ do
11         for $i = 1 : N$ do
12             if Terminal-$i$ senses the channel is idle and
                 DQN of terminal-$i$ outputs transmit then
13                 Terminal-$i$ transmits with probability
                     $p_{\mathsf{tx}}$ in this time slot.
14             else
15                 Terminal-$i$ stays silent in this time slot.
16     $R_k$ = average utility over time $T$.
17 if $R_1 < R_2$ then
18     $m_{\mathsf{min}} = m_1$, $m_1 = m_2$, $m_2 = \frac{\sqrt{5}-1}{2} m_1 + \frac{3-\sqrt{5}}{2} m_{\mathsf{max}}$,
19 else
20     $m_{\mathsf{max}} = m_2$, $m_2 = m_1$, $m_1 = \frac{\sqrt{5}-1}{2} m_2 + \frac{3-\sqrt{5}}{2} m_{\mathsf{min}}$
21 if $|m_{\mathsf{max}} - m_{\mathsf{min}}| > \epsilon$ then
22     Return to Step 4.
23 else
24     TT training is completed.
```

B. MARL Schemes for Comparisons

We also leverage the state-of-the-art MARL algorithms in this setting and investigate their performances. In this subsection, two such algorithms are introduced which achieve reasonably good performances, whereas under-performing compared with TT.

MARL with Auxiliary State: One of the main technique in MARL to combat the instability of multi-agent learning is to use auxiliary state to let each terminal have knowledge of other terminals' states. In this way, instead of treating the environment stationary using conventional RL methods while the real environment involves the interplay with other terminals and thus non-stationary, leveraging auxiliary state can explicitly model the behavior of other agents.

Specifically, we adopt a Deep Deterministic Policy Gradient (DDPG) [21] based architecture for the agents. Each agent only has information about its own state, the auxiliary state indicates whether its last transmission attempt was successful or not (i.e., a collision flag). The action it takes represents the transmission probability for the agent at the moment, and is a real value within the $0-1$ interval. The reward it receives is provided by the FC, which computes it based on a global fairness formula—for the homogeneous AoI optimization case, the sum of squared AoIs observed at the FC is used. Note that this reward is received only if the agent manages to transmit successfully; if the transmission fails, the reward it receives is 0. All agents are trained simultaneously. Although this seems more efficient compared with TT, we find that, in general, it is quite difficult to achieve both stability and scalability. As shown later in the simulation results, performance degradation is evident based on this approach.

Cross-Entropy Based Method: Another simple, yet sometimes powerful, scheme is also tested, namely the cross-entropy (CE) based method. The CE based method can be categorized as an evolutionary algorithm—it selects the parameters that survive the natural selection, i.e., perform better than others, in each iteration and evolves over time. The implementation details are omitted in this paper, as they have been mentioned in our previous work [1].

IV. Simulation Results

The performance of the proposed schemes is investigated by computer simulations. First, the time-average AoI of CE-based approach, MARL with auxiliary state and TT schemes is tested in comparison with the genie-aided Whittle's index approach which schedules the terminal with the largest index. The packet arrival rate is 0.1 packets/ms, the length of a minislot is 0.01 ms, and the numbers of agents are 3 and 30 in Fig. 2(a) and 2(b) respectively, and the y-axis is the running average AoI over time (x-axis). It is observed that while the CE-based approach can converge slowly when there are 3 agents, it fails when scaling up to 30 agents; the same scalability issue also occurs with the MARL approach with auxiliary state—it takes too long to converge when the number of agents scale up to, e.g., 100 agents. In the mean time, it is found that, since the training for each agent can be made

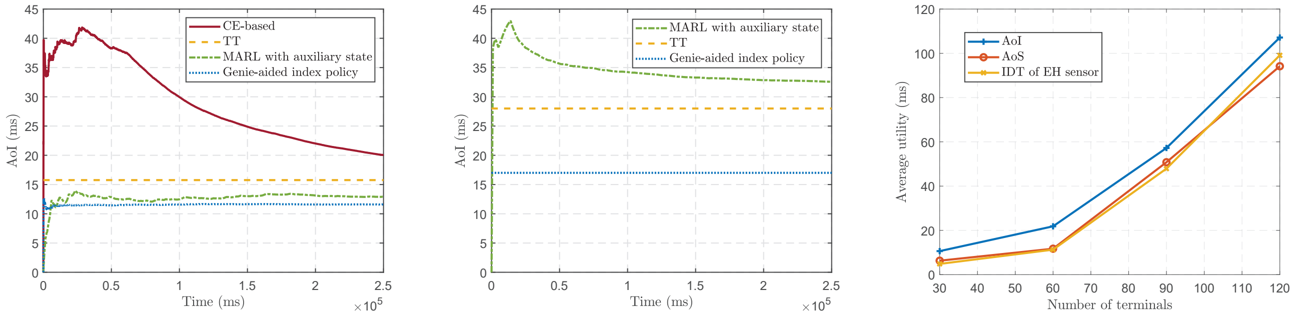

Fig. 2. Performance evaluations for (a) AoI with 3 agents; (b) AoI with 30 agents; (c) Diversified QoS requirements for more agents. The packet arrival rates are 0.1 packets/ms, the energy arrival rate is 0.2 packets/ms, energy buffer size is 1, and the length of a minislot is 0.01 ms.

offline by storing a mapping table from transmission tax m to the model parameters $\boldsymbol{w}_i$, the time for TT to converge only consists of the time for the golden search iteration, which is independent of the number of agents. Therefore, TT has good scalability thanks to its separate training architecture.

When scaling up to more agents in Fig. 2(c), the average utility—which in this case include AoI, AoS and energy harvesting (EH) sensors that are optimizing the IDT and each occupy one third of the total number of terminals—shows that the proposed framework is scalable and applicable for heterogeneous QoS requirements. The energy arrival rate is 0.2 packets/ms and the energy buffer size is one for the EH sensor. The detailed description of the utilities is omitted due to lack of space.

V. Conclusions

In this paper, we propose a scalable distributed learning based framework for SMART. By arming the terminals with situationally-awareness, only valuable packets—from the perspective of specific applications with diversified QoS requirements, e.g., AoI, IDT and AoS—are transmitted in the wireless uplinks to improve the efficiency and latency performance of current 5G networks. By introducing the transmission tax thus decoupling the training processes of terminals, the proposed framework, as revealed by the simulation results, is scalable and stable. Future work includes more realistic experiments and considerations for specific applications.

Acknowledgement

This work is sponsored in part by the Nature Science Foundation of China (No. 61861136003, No. 91638204, No. 61571265, No. 61621091) and Hitachi Ltd.

References

[1] Z. Jiang, S. Zhou, and Z. Niu, "Distributed policy learning based random access for diversified QoS requirements," in *IEEE International Conference on Communications (ICC)*, Jun 2019.

[2] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *IEEE INFOCOM*, Apr 2018, pp. 1970–1978.

[3] 3GPP TR. 21.915 v 0.6.0 "Initial access and mobility".

[4] J. Zhang, L. Lu, Y. Sun, Y. Chen, J. Liang, J. Liu, H. Yang, S. Xing, Y. Wu, J. Ma, I. B. F. Murias, and F. J. L. Hernando, "PoC of SCMA-based uplink grant-free transmission in UCNC for 5G," *IEEE J. Select. Areas Commun.*, vol. 35, no. 6, pp. 1353–1362, Jun 2017.

[5] J. Cheng, C. Lee, and T. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," in *2011 IEEE GLOBECOM Workshops (GC Wkshps)*, Dec 2011, pp. 368–372.

[6] R. Cheng, J. Chen, D. Chen, and C. Wei, "Modeling and analysis of an extended access barring algorithm for machine-type communications in LTE-A networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2956–2968, Jun 2015.

[7] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun 2017.

[8] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *IEEE INFOCOM*, Mar 2012, pp. 2731–2735.

[9] J. Zhong, R. D. Yates, and E. Soljanin, "Two freshness metrics for local cache refresh," in *IEEE Int'l Symp. Info. Theory*, Jun 2018, pp. 1924–1928.

[10] B. Yin, S. Zhang, Y. Cheng, L. X. Cai, Z. Jiang, S. Zhou, and Z. Niu, "Only those requested count: Proactive scheduling policies for minimizing effective age-of-information," in *IEEE INFOCOM*, April 2019.

[11] X. Guo, R. Singh, P. R. Kumar, and Z. Niu, "A risk-sensitive approach for packet inter-delivery time optimization in networked cyber-physical systems," *IEEE/ACM Trans. Netw.*, vol. 26, no. 4, pp. 1976–1989, Aug. 2018.

[12] Z. Jiang, B. Krishnamachari, S. Zhou, and Z. Niu, "Can decentralized status update achieve universally near-optimal age-of-information in wireless multiaccess channels?" in *International Teletraffic Congress (ITC 30)*, Sep 2018.

[13] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Niu, "Decentralized status update for age-of-information optimization in wireless multiaccess channels," in *IEEE Int'l Symp. Info. Theory*, 2018.

[14] ——, "Timely status update in wireless uplinks: Analytical solutions with asymptotic optimality," *IEEE Internet of Things Journal*, 2018.

[15] Z. Jiang, S. Zhou, Z. Niu, and Y. Cheng, "A unified sampling and scheduling approach for status update in wireless multiaccess networks," in *IEEE INFOCOM*, April 2019, pp. 1–9.

[16] Y. Gai, S. Ganesan, and B. Krishnamachari, "The saturation throughput region of p-persistent CSMA," in *Information Theory and Applications Workshop*, Feb 2011, pp. 1–4.

[17] J. Ni, B. Tan, and R. Srikant, "Q-CSMA: Queue-length-based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 825–836, Jun 2012.

[18] L. Peshkin, K.-E. Kim, N. Meuleau, and L. P. Kaelbling, "Learning to cooperate via policy search," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2000, pp. 489–496.

[19] D. P. Bertsekas, *Dynamic programming and optimal control.* Athena scientific Belmont, MA, 1995, vol. 1.

[20] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[21] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *ICML*, 2014.