Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks

Arjun Anand*, Gustavo de Veciana*, and Sanjay Shakkottai*

*Department of Electrical and Computer Engineering, The University of Texas at Austin

***Abstract*—Emerging 5G systems will need to efficiently support both enhanced mobile broadband traffic (eMBB) and ultra-low-latency communications (URLLC) traffic. In these systems, time is divided into slots which are further sub-divided into minislots. From a scheduling perspective, eMBB resource allocations occur at slot boundaries, whereas to reduce latency URLLC traffic is pre-emptively overlapped at the minislot timescale, resulting in selective superposition/puncturing of eMBB allocations. This approach enables minimal URLLC latency at a potential rate loss to eMBB traffic.**

We study joint eMBB and URLLC schedulers for such systems, with the dual objectives of maximizing utility for eMBB traffic while immediately satisfying URLLC demands. For a linear rate loss model (loss to eMBB is linear in the amount of URLLC superposition/puncturing), we derive an optimal joint scheduler. Somewhat counter-intuitively, our results show that our dual objectives can be met by an iterative gradient scheduler for eMBB traffic that anticipates the expected loss from URLLC traffic, along with an URLLC demand scheduler that is oblivious to eMBB channel states, utility functions and allocation decisions of the eMBB scheduler. Next we consider a more general class of (convex/threshold) loss models and study optimal online joint eMBB/URLLC schedulers within the broad class of channel state dependent but minislot-homogeneous policies. A key observation is that unlike the linear rate loss model, for the convex and threshold rate loss models, optimal eMBB and URLLC scheduling decisions do not de-couple and joint optimization is necessary to satisfy the dual objectives. We validate the characteristics and benefits of our schedulers via simulation.

***Index Terms*—wireless scheduling, URLLC traffic, 5G systems**

I. Introduction

An important requirement for 5G wireless systems is its ability to efficiently support both broadband and ultra reliable low-latency communications. On one hand enhanced Mobile Broadband (eMBB) might require gigabit per second data rates (based on a bandwidth of several 100 MHz) and a moderate latency (a few milliseconds). On the other hand, Ultra Reliable Low Latency Communication (URLLC) traffic requires extremely low delays (0.25-0.3 msec/packet) with very high reliability (99.999%) [1]. To satisfy these heterogenous requirements, the 3GPP standards body has proposed an innovative *superposition/puncturing* framework for multiplexing URLLC and eMBB traffic in 5G cellular systems[1].

The proposed scheduling framework has the following structure [1]. As with current cellular systems, time is divided into slots, with a proposed one millisecond (msec) slot duration. Within each slot, eMBB traffic can share the bandwidth over the time-frequency plane (see Figure 1). The sharing mechanism can be opportunistic (based on the channel states of various users); however, the eMBB shares are decided by the beginning, and fixed for the duration of a slot[2]. Further the new framework also allows aggregation of eMBB slots where transmissions to an eMBB user over consecutive slots are coded together to achieve better coding gains resulting from long codewords while reducing overheads due to control signals. This results in better spectral efficiency as compared to the OFDMA frame structure of LTE [3].

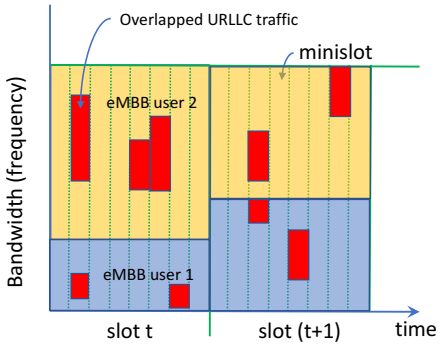

Fig. 1. Illustration of superposition/puncturing approach for multiplexing eMBB and URLLC: Time is divided into slots, and further subdivided into minislots. eMBB traffic is scheduled at the beginning of slots (sharing frequency across two eMBB users), whereas URLLC traffic can be dynamically overlapped (superpose/puncture) at any minislot.

URLLC downlink packets may arrive during an ongoing eMBB transmission; if tight latency constraints are to be satisfied, they cannot be queued until the next slot. Instead each eMBB slot is divided into minislots, each of which has a 0.125 msec duration[3]. Thus upon arrival URLLC packets can be immediately scheduled in the next minislot *on top of the ongoing eMBB transmissions*. If the Base Station (BS) chooses non-zero transmission powers for both eMBB and overlapping URLLC traffic, then this is referred to as

[1]An earlier version of this work appears in the Proceedings of IEEE Infocom 2018, Honolulu, HI, [2].

[2]The sharing granularity among various eMBB users is at the level of Resource Blocks (RB), which are small time-frequency rectangles within a slot. In LTE today, these are (1 msec $\times$ 180 KHz), and could be smaller for 5G systems.

[3]In 3GPP, the formal term for a 'slot' is eMBB TTI, and a 'minislot' is a URLLC TTI, where TTI expands to Transmit Time Interval.

superposition. If eMBB transmissions are allocated zero power when URLLC traffic is overlapped, then it is referred to as *puncturing* of eMBB transmissions. To achieve high reliability URLLC transmissions are by design protected through coding and HARQ if necessary. At the end of an eMBB slot, the BS can signal eMBB users the locations, if any, of URLLC superposition/puncturing. eMBB users can then use this information to decode transmissions, with some possible loss of rate depending on the amount of URLLC overlap. See [1], [4] for additional details.

A key problem in this setting is the *joint scheduling of eMBB and URLLC traffic over two time-scales.* At the slot boundary, resources are allocated to eMBB users (with possible aggregation of slots) based on their channel states and utilities, in effect, allocating long term rates to optimize high-level goals (e.g. utility optimization). Meanwhile, at each minislot boundary, the (stochastic) URLLC demands are placed onto previously scheduled and ongoing eMBB transmissions. Decisions on the placement of such overlaps across scheduled eMBB user(s) will impact the rates they will see on that slot. Thus we have a coupled problem of jointly optimizing the scheduling of eMBB users on slots with the placement of URLLC demands across minislots.

A. Main Contributions

This paper is, to our knowledge, the first to formalize and solve the joint eMBB/URLLC scheduling problem described above. We consider various models for the eMBB rate loss associated with URLLC superposition/puncturing, for which we characterize the associated feasible throughput regions and propose online joint scheduling algorithms as detailed below.

Linear Model: When the rate loss to eMBB is directly proportional to the fraction of superposed/punctured minislots, we show that the joint optimal scheduler has a nice decomposition. Despite having non-linear utility functions and time-varying channel states, the stochastic URLLC traffic can be *uniform-randomly placed* in each minislot, while the eMBB scheduler can be scheduled via a greedy iterative gradient algorithm that only accounts for the *expected* rate loss due to the URLLC traffic.

Convex Model: For more general settings where the rate loss can be modeled by a convex function, the solution does not have the decomposition property as in the linear model and hence, the finding the optimal solution is challenging. Therefore, we restrict to a simpler class of joint scheduling policies called as *minislot-homogeneous* joint scheduling policies where the URLLC placement policy does not change across the minislots in an eMBB slot. In this setting, we characterize the capacity region and derive concavity conditions under which we can derive the effective rate seen by eMBB users (post-puncturing by URLLC traffic). We then develop a stochastic approximation algorithm which jointly schedules eMBB and URLLC traffic, and show that it asymptotically maximizes the utility for eMBB users while satisfying URLLC demands. We also show that for convex functions which are *homogeneous*, minislot-homogeneous joint scheduling policies are optimal within the larger class of *causal* and *non-anticipative* joint scheduling policies. Further for the convex loss model, we show that it is better to schedule eMBB users to share bandwidth (i.e. slice across frequency, see also Fig. 4), and let each user occupy the entire slot duration to mitigate rate loss due to URLLC puncturing.

Threshold Model: Finally we consider a loss model, where eMBB traffic is unaffected by puncturing until a threshold is reached; beyond this threshold it suffers complete throughput loss (a 0-1 rate loss model). We consider two broad classes of minislot homogeneous policies, where the URLLC traffic is placed in minislots in proportion to the eMBB resource allocations (Rate Proportional (RP)) or eMBB loss thresholds (Threshold Proportional (TP)). We motivate these policies (e.g. TP minimizes the probability of any eMBB loss in an eMBB slot) and derive the associated throughput regions. Finally, we utilize the additional structure underlying the RP and TP Placement policies along with the shape of the threshold loss function to derive fast gradient algorithms that converge and provably maximize utility.

B. Related Work

Resource allocation, utility maximization and opportunistic scheduling for downlink wireless systems have been intensely studied in the last two decades, and have had a major impact on cellular standards. We refer to [5], [6] for a survey of the key results. In this paper, we focus on joint scheduling of URLLC and eMBB traffic. From an application point of view, there have been several studies arguing for the need to support URLLC services (e.g. for industrial automation) [7], [8], [9].

With demand of both broadband and low-latency services growing, there has been rapid developments in the 5G standardization efforts in 3GPP. Of key relevance to this paper, the 3GPP RAN WG1 has focused on standardizing slot structure for eMBB and URLLC, and have been evaluating signaling and control channels to support superposition and puncturing in recent meetings [1], [4]. We specifically refer the reader to Sections 8.1.1.3.4 – 8.1.1.3.6 in [4] for current proposals.

Beyond standards, recent work has focused on system level design for such systems (overheads, packet sizes, control channel structure, etc.) [3], [10], [11]. Of particular note, [10] argues (based on system level simulation and queuing models) that statically partitioning bandwidth between eMBB and URLLC is very inefficient. There have also been several studies focusing on physical layer aspects of URLLC (coding and modulation, fading, link budget) [12], [13].

Efficient sharing of radio resources between eMBB and URLLC traffic has been discussed in literature, see [14], [15], [16]. In [14], the authors have considered joint optimization of resource allocation for eMBB and URLLC traffic. However, they do not use puncturing/superposition mechanisms to share resources. Some works ([15], [16]) use information theoretic results to obtain expressions for the average eMBB rates under URLLC puncturing for various decoding schemes for uplink eMBB traffic punctured/superposed by URLLC users. However, they do not consider the design of joint scheduling

algorithms for eMBB and URLLC traffic. To the best of our knowledge, our paper is the first to explore the resource allocation issues for joint scheduling of URLLC and eMBB traffic using puncturing/superposition based mechanisms.

II. System Model

Traffic model: We consider a wireless system supporting a fixed set $\mathcal{U}$ of backlogged eMBB users and a stationary process of URLLC demands. eMBB scheduling decisions are made across slots while URLLC demands arrive and are immediately scheduled in the next minislot. In this section we shall consider the case where eMBB all users receive resources for slots without using slot aggregation even though more flexible resource allocations which can possibly include slot aggregation and splitting are proposed in 5G standards [3]. We shall justify this choice in Sec. IV-E. Each eMBB slot has an associated set of minislots where the set $\mathcal{M} = \{1, \dots |\mathcal{M}|\}$ denotes their indices. URLLC demands across minislots are modeled as an independent and identically distributed (i.i.d.) random process. We let the random variables $(D(m), m \in \mathcal{M})$ denote the URLLC demands per minislot for a typical eMBB slot and let D be a random variable whose distribution is that of the aggregate URLLC demand per eMBB slot, i.e., $D \sim \sum_{m\in\mathcal{M}} D(m)$ with, cumulative distribution function $F_D(\cdot)$ and mean $E[D] = \rho$. We assume demands have been normalized so the maximum URLLC demand per minislot is f and the maximum aggregate demands per eMBB slot is $f \times |\mathcal{M}| = 1$ i.e., all the frequency-time resources are occupied. URLLC demands per minislot exceeding the system capacity are blocked by URLLC scheduler thus $D \leq 1$ almost surely. The system is engineered so that blocked URLLC traffic on a minislot is a rare event, i.e., satisfies the desired reliability on such traffic.

Wireless channel variations: The wireless system experiences channel variations each eMBB slot which are modeled as an i.i.d. random process over a set of channel states $\mathcal{S} = \{1, \dots, |\mathcal{S}|\}$. Let S be a random variable modeling the distribution over the states in a typical eMBB slot with probability mass function $p_S(s) = P(S = s)$ for $s \in \mathcal{S}$. For each channel state s eMBB user u has a known peak rate $\hat{r}_u^s$. The wireless system can choose what proportions of the frequency-time resources to allocate to each eMBB user on each minislot for each channel state. This is modeled by a matrix $\boldsymbol{\phi} \in \Sigma$ where

$$\Sigma := \Big\{ \boldsymbol{\phi} \in \mathbb{R}_+^{|\mathcal{U}|\times|\mathcal{M}|\times|\mathcal{S}|} \;\Big|\; \sum_{u\in\mathcal{U}} \phi_{u,m}^s = f, \forall m \in \mathcal{M}, s \in \mathcal{S} \Big\} \quad (1)$$

and where the element $\phi_{u,m}^s$ represents the fraction of resources allocated to user u in mini slot m in channel state s. We also let $\phi_u^s = \sum_{m\in\mathcal{M}} \phi_{u,m}^s$, i.e., the total resources allocated to user u in an eMBB slot in channel state s. Now assuming no superposition/puncturing if the system is in channel state s and the eMBB scheduler chooses an allocation $\boldsymbol{\phi}$ the rate r_u allocated to user u would be given by $r_u = \phi_u^s \hat{r}_u^s$. The scheduler is assumed to know the channel state and can thus opportunistically exploit such variations in allocating resources to eMBB users. Note that for simplicity, we adopt a flat-fading model, namely, the rate achieved by an user is directly proportional to the fraction of bandwidth allocated to it (the scaling factor is the peak rate of the user for the current channel state).

Class of joint eMBB/URLLC schedulers: We consider a class of stationary joint eMBB/URLLC schedulers denoted by Π satisfying the following properties. A scheduling policy combines a possibly state dependent eMBB *resource allocation* matrix $\boldsymbol{\phi}$ per slot with a URLLC *demand placement* strategy across minislots. The placement strategy may impact the eMBB users' rates since it affects the URLLC superposition/puncturing loads they will experience. As mentioned earlier in discussing the traffic model, in order to meet low latency requirements URLLC traffic demands are scheduled immediately upon arrival or blocked. The scheduler is assumed to be *causal* so it only knows the current (and past) channel states and peak rates $\hat{r}_u^s$ for all $u \in \mathcal{U}$ and $s \in \mathcal{S}$ but does not know the realization of future channels or URLLC traffic demands. In making superposition/puncturing decisions across minislots, the scheduler can use knowledge of the previous placement decisions that were made. In addition the scheduler is assumed to know (or able measure over time) the channel state distribution across eMBB slots and URLLC demand distributions per minislot i.e., that of $D(m)$, and per eMBB slot, i.e., D, and thus in particular knows $\rho = E[D]$.

In summary a joint scheduling policy $\pi \in \Pi$ is thus characterized by the following:

- an eMBB resource allocation $\boldsymbol{\phi}^\pi \in \Sigma$ where $\phi_{u,m}^{\pi,s}$ denotes the fraction of frequency-time slot resources allocated to eMBB user u on minislot m when the system is in state s.
- the distributions of URLLC loads across eMBB resources induced by its URLLC placement strategy, denoted by random variables $\mathbf{L}^\pi = (L_{u,m}^{\pi,s} | u \in \mathcal{U}, m \in \mathcal{M}, s \in \mathcal{S})$ where $L_{u,m}^{\pi,s}$ denotes the URLLC load superposed/puncturing the resource allocation of user u on minislot m when the channel is in state s.

The distributions of $L_{u,m}^{\pi,s}$ and their associated means $\bar{l}_{u,m}^{\pi,s}$ depend on the joint scheduling policy π, but for all states, users and minislots satisfy

$$L_{u,m}^{\pi,s} \leq \phi_{u,m}^{\pi,s} \quad \text{almost surely.}$$

In the sequel we let $L_u^{\pi,s} = \sum_{m\in\mathcal{M}} L_{u,m}^{\pi,s}$, i.e., the aggregate URLLC traffic superposed/puncturing user u in channel state s, and denote its mean by $\bar{l}_u^{\pi,s}$ and note that

$$L_u^{\pi,s} \leq \phi_u^{\pi,s} \quad \text{almost surely.}$$

We also let $L^{\pi,s} := \sum_{u\in\mathcal{U}} L_u^{\pi,s}$ denote the aggregate induced load and note that any policy π and for any state s we have that

$$\rho = \mathrm{E}[D] = \mathrm{E}[L^{\pi,s}] = \mathrm{E}[\sum_{u\in\mathcal{U}} L_u^{\pi,s}] = \sum_{u\in\mathcal{U}} \bar{l}_u^{\pi,s}.$$

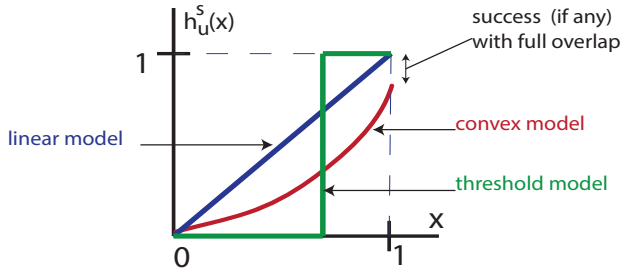

Fig. 2. The illustration exhibits the rate loss function for the various models considered in this paper, linear, convex and threshold.

Modeling superposition/puncturing and eMBB capacity regions: Under a joint scheduling policy π we model the rate achieved by an eMBB user u in channel state s by a random variable

$$R_u^{\pi,s} \quad = \quad f_u^s(\phi_u^{\pi,s}, L_u^{\pi,s}), \tag{2}$$

where the *rate allocation function* $f_u^s(\cdot,\cdot)$ models the impact of URLLC superposition/puncturing – one would expect it to be increasing in the first argument (the allocated resources) and decreasing in the second argument (the amount superposition/puncturing by URLLC traffic). Under our system model we have that

$$R_u^{\pi,s} \le f_u^s(\phi_u^{\pi,s}, 0) = \phi_u^{\pi,s}\hat{r}_u^s \text{ almost surely,}$$

with equality if there is no superposition/puncturing, i.e., when $l_u^s = 0$. Let $\overline{r}_u^{\pi,s} = E[R_u^{\pi,s}]$ denote the mean rates achieved by user u in state s under the URLLC superposition/puncturing distribution induced by scheduling policy π.

Models for Throughput Loss: In the sequel we shall consider specific forms of superposition/puncturing loss models: *(i)* linear, *(ii)* convex, and *(iii)* threshold models.

We rewrite the rate allocation function in (2) as the difference between the peak throughput and the loss due to URLLC traffic, and consider functions that can be decomposed as:

$$f_u^s(\phi_u^s, l_u^s) = \hat{r}_u^s\phi_u^s\left(1 - h_u^s\left(\frac{L_u^{\pi,s}}{\phi_u^s}\right)\right),$$

where $h_u^s : [0,1] \to [0,1]$ is the *rate loss function* and captures the relative rate loss due to URLLC overlap on eMBB allocations. The puncturing models we study now map directly to structural assumptions on the rate loss function $h_u^s(\cdot)$; namely it is a non-decreasing function, and is one of *linear, convex, or threshold* as shown in Figure 2.

Linear Model: Under the linear model, the expected rate for user u in channel state s for policy π is given by

$$r_u^{\pi,s} = \mathrm{E}[f_u^s(\phi_u^{\pi,s}, L_u^{\pi,s})] = \hat{r}_u^s(\phi_u^{\pi,s} - \bar{l}_u^{\pi,s}),$$

i.e., $h_u^s(x) = x$, and the resulting rate to eMBB users is a linear function of both the allocated resources and mean induced URLLC loads. This model is motivated by basic results for the channel capacity of AWGN channel with erasures, see [17] for more details. Our system in a given network state can be approximated as an AWGN channel with erasures, when the slot sizes are long enough so that the physical layer error control coding of eMBB users use long code-words. Further, there is a dedicated control channel through which the scheduler can signal to the eMBB receiver indicating the positions of URLLC overlap. Indeed such a control channel has been proposed in the 3GPP standards [1]. Note that under this model the rate achieved by a given user depends on the aggregate superposition/puncturing it experiences, i.e., does not depend on which minislots and frequency bands it occurs. We discuss scheduling policies for linear loss models in Section III.

Convex Model: In the convex model, the rate loss function $h_u^s(\cdot)$ is convex (see Figure 2), and the resulting rate for eMBB user u in channel state s under policy π is given by

$$r_u^{\pi,s} = \mathrm{E}[f_u^s(\phi_u^{\pi,s}, L_u^{\pi,s})] = \hat{r}_u^s\phi_u^{\pi,s}\left(1 - E\left[h_u^s\left(\frac{L_u^{\pi,s}}{\phi_u^{\pi,s}}\right)\right]\right).$$

This covers a broad class of models, and is discussed in Section IV.

Threshold Model: Finally the threshold model is designed to capture a simplified packet transmission and decoding process in an eMBB receiver. The data is either received perfectly or it is lost depending on the amount of superposition/puncturing. With slight abuse of notation we shall let h_u^s also depend on both the relative URLLC load and the eMBB user allocation, i.e., $h_u^s(x) = \mathbf{1}(x \ge t_u^s(\phi_u^s))$ where the threshold in turn is an increasing function $t_u^s(\cdot)$ satisfying $x \ge t_u^s(x) \ge 0$. Such thresholds might reflect various engineering choices where codes are adapted when users are allocated more resources, so as to be more robust to interference/URLLC superposition/puncturing. The resulting rate for eMBB user u in channel state s and policy π is then given by

$$r_u^{\pi,s} = \hat{r}_u^s\phi_u^{\pi,s}P(L_u^{\pi,s} \le \phi_u^{\pi,s}t_u^s(\phi_u^{\pi,s})).$$

While such a sharp falloff is somewhat extreme, it is nevertheless useful for modeling short codes that are designed to tolerate a limited amount of interference. In practice one might expect a smoother fall off, perhaps more akin to the convex model, e.g., when hybrid ARQ (HARQ) is used. We discuss polices under the threshold based model in Section V.

Capacity set for eMBB traffic: We define the capacity set $\mathcal{C} \subset \mathbb{R}_+^{|\mathcal{U}|}$ for eMBB traffic as the set of long term rates achievable under policies in Π. Let $\mathbf{c}^\pi = (c_u^\pi | u \in \mathcal{U})$ where

$$c_u^\pi = \sum_{s\in\mathcal{S}} r_u^{\pi,s}p_S(s).$$

Then the capacity is given by

$$\mathcal{C} = \{\mathbf{c} \in \mathbb{R}_+^{|\mathcal{U}|} \mid \exists\ \pi \in \Pi \text{ such that } \mathbf{c} \le \mathbf{c}^\pi\}.$$

Note that this capacity region depends on the scheduling policies under consideration as well as the distributions of the channel states and URLLC demands.

Scheduling objective: URLLC priority and eMBB utility maximization: As mentioned earlier, URLLC traffic is immediately scheduled upon arrival, in the next minislot, i.e,

no queuing is allowed. Thus if demands exceed the system capacity on a given minislot traffic would be lost. However, we assume that the system has been engineered so that such URLLC overloads are extremely rare, and thus URLLC traffic can meet extremely low latency requirements with high reliability[4]. For eMBB traffic we adopt a utility maximization framework wherein each eMBB user u has an associated utility function $U_u(\cdot)$ which is a strictly concave, continuous and differentiable of the average rate c_u^π experienced by the user. Our aim is to characterize optimal rate allocations associated with the utility maximization problem:

$$\max_{\mathbf{c}}\{\sum_{u\in\mathcal{U}} U_u\left(c_u\right) \mid \mathbf{c}\in\mathcal{C}\}, \tag{3}$$

and determine a scheduling policy π that will realize such allocations.

III. Linear Model for Superposition/Puncturing

In any state s, the optimal joint eMBB/URLLC scheduler may either 1) protect the user with the lower channel rate by placing less URLLC traffic into its frequency resources to ensure fairness or 2) opportunistically place URLLC traffic so that the user with a better channel gets a higher rate to improve the overall system throughput. The solution for any state is complex function of network states and their distribution and user utility functions and in general, eMBB scheduling and URLLC puncturing may be dependent. In this section, we show a surprising result – despite having non-linear utility functions, if the loss functions are linear and the eMBB scheduler is intelligent (i.e., takes into the degradation of rates due to puncturing), then the URLLC scheduler can be *oblivious to the channel states, utility functions and the actual rate allocations of the eMBB scheduler.*

A. Characterization of capacity region

Let us consider the capacity region for a wireless system based on linear superposition/puncturing model under a restricted class of policies Π^{LR} that combine feasible eMBB allocations $\boldsymbol{\phi}\in\Sigma$ with random placement of URLLC demands uniformly over the bandwidth across minislots. Note that the notation LR stands for linear loss model (L) with random (R) placement of URLLC traffic. For any $\pi\in\Pi^{LR}$ with eMBB allocation $\boldsymbol{\phi}^\pi$ the mean induced loads under such randomization for each state $s\in\mathcal{S}$ and minislot $m\in\mathcal{M}$ will satisfy $\bar{l}_{u,m}^{\pi,s} = \rho\phi_{u,m}^{\pi,s}$. Indeed randomization clearly leads to an induced loads that are proportional to the eMBB allocations on a per mini-slot basis, but also per eMBB slot, i.e., $\bar{l}_u^{\pi,s} = \rho\phi_u^{\pi,s}$. Thus for our linear loss model we have that

$$r_u^{\pi,s} = \hat{r}_u^s(\phi_u^{\pi,s} - \bar{l}_u^{\pi,s}) = \hat{r}_u^s\phi_u^{\pi,s}(1-\rho).$$

Hence the overall user rates achieved under such a policy are given by $\mathbf{c}^\pi = (c_u^\pi | u\in\mathcal{U})$ where

$$c_u^\pi = \sum_{s\in\mathcal{S}} \hat{r}_u^s\phi_u^{\pi,s}(1-\rho)p_S(s).$$

[4]Note that since we allow URLLC traffic in the entire system bandwidth, such overload events are very rare.

The capacity region associated with policies that use URLLC uniformly randomized placement is thus given by

$$\begin{aligned}\mathcal{C}^{LR} &= \{\mathbf{c}\in\mathbb{R}_+^{|\mathcal{U}|} \mid \exists\pi\in\Pi^{LR} \text{ s.t. } \mathbf{c}\le\mathbf{c}^\pi\}\\ &= \{\mathbf{c}\in\mathbb{R}_+^{|\mathcal{U}|} \mid \exists\boldsymbol{\phi}\in\Sigma \text{ s.t. } \mathbf{c}\le\mathbf{c}^{\boldsymbol{\phi}}\},\end{aligned}$$

where we have abused notation by using $\mathbf{c}^{\boldsymbol{\phi}}$ to represent the throughput achieved under policy π that uses eMBB resource allocation $\boldsymbol{\phi}$ and uniformly randomized URLLC demand placement. Finally note that for any fixed $\rho\in(0,1)$, $\mathcal{C}^{LR}$ is a closed and bounded convex region. This is because an affine map of a convex region remains convex; hence multiplying the constraints on the capacity region defined by $\boldsymbol{\phi}$ by a constant $(1-\rho)$ preserves convexity of the rate region.

Theorem 1. *For a wireless system under the linear superposition/puncturing loss model we have that $\mathcal{C}=\mathcal{C}^{LR}$.*

The proof is deferred to the Appendix A. In other words the throughput $\mathbf{c}^\pi\in\mathcal{C}$ achieved by any feasible policy $\pi\in\Pi$ can also be achieved by policy π', with a possibly different eMBB resource allocation policy than π but utilizing uniform random placement of URLLC demands across mini-slots.

B. Utility maximizing joint scheduling

Given the result in Theorem 1 we now restate the utility maximization problem as optimizing solely over joint scheduling policies that use URLLC random placement policies, as follows:

$$\begin{aligned}\max_{\boldsymbol{\phi}\in\Sigma} \quad & \sum_{u\in\mathcal{U}} U_u(c_u^{\boldsymbol{\phi}}),\\ \text{s.t.} \quad & c_u^{\boldsymbol{\phi}} = \sum_{s\in\mathcal{S}} \hat{r}_u^s\phi_u^s(1-\rho)p_S(s), \quad \forall u\in\mathcal{U}.\end{aligned}$$

The above optimization problem has a strictly concave cost function and convex constraints. Thus, at face-value, it appears that we can apply the gradient scheduler introduced in [18], which is an online algorithm designed to converge to the solution of similar optimization problem. This observation is approximately correct, but subject to two modifications.

First, the setting in [18] has deterministic rates in each channel state. However, in our case, in each channel state, the rates are stochastic due to puncturing by URLLC traffic (this results in the $(1-\rho)$ correction). This can be easily addressed by modifying the setting in [18]; the finite state and i.i.d. nature of puncturing implies that the proofs in [18] hold with minor modifications; we skip the details.

The second issue is somewhat more nuanced. In current wireless systems (e.g. LTE) and proposals for 5G systems, a slot is partitioned into a collection of Resource Blocks (RB), where each RB is a time-frequency rectangle (1 msec $\times$ 180 KHz in LTE). Importantly, these RBs can be individually allocated to different eMBB users. If we now apply the gradient scheduler in [18] to our setting, the result will be that all RBs in a slot will be allocated to the same user. While this is no-doubt asymptotically optimal, it seems intuitive that sharing RBs across users even within a slot will lead to better short-term performance. Indeed this intuition has been explored

in the context of iterative MaxWeight algorithms to provide formal guarantees, see [19], [20]. The high level idea is that even within a slot, RB allocations are done iteratively, where future RB allocations need to account for prior rate allocations even within the same slot. This is formalized below, where we describe our proposed joint eMBB-URLLC scheduler.

The URLLC scheduler: As explained in the previous section, the URLLC scheduler places the URLLC traffic uniformly at random in each minislot.

The eMBB scheduler: Let there be B resource blocks available for allocation every eMBB slot, indexed by $1, 2, \ldots, B$. Let $\overline{R}_u(t-1)$ be the random variable denoting the average rate received by eMBB user up to eMBB slot $t-1$. Let $\overline{r}_u(t-1)$ be a realization of $\overline{R}_u(t-1)$. In any eMBB slot t we schedule an user $u(b)$ in RB b such that

$$u(b) \in \text{argmax} \left\{ \hat{r}_u^s U_u^{'} \left(\overline{r}_u^{\epsilon} \left(b-1, t \right) \right), u = 1, 2, \ldots, \mathcal{U} \right\}, \quad (4)$$

where $\overline{r}_u^{\epsilon}\left(b-1,t\right)$ is an *estimate* of the average rate received by eMBB user u till slot t which is iteratively updated as follows:

$$\overline{r}_u^{\epsilon}\left(b,t\right) = \begin{cases} \overline{r}_u(t-1), & b = 0, \\ (1-\epsilon)\,\overline{r}_u^{\epsilon}\left(b-1,t\right) & \\ +\epsilon\left(\hat{r}_u^s \frac{1}{B}(1-\rho)\mathbb{1}\left(i = u(b)\right)\right), & b \neq 0. \end{cases} \quad (5)$$

In the above equation, ϵ is a small positive value. At the end of eMBB slot t, the eMBB scheduler receives feedback from the eMBB receivers indicating the actual rates received by the eMBB users due to allocations. We denote the rate received eMBB user u in slot by the random variable $R_u(t)$ and its realization by $r_u(t)$. We finally update $\overline{r}_u(t)$ as follows:

$$\overline{r}_u(t) = (1-\epsilon)\,\overline{r}_u(t-1) + \epsilon r_u(t). \quad (6)$$

This scheduler and update equations are analogous to the gradient algorithm [18] (see also iterative algorithms in [19], [20]). The optimality proof of this algorithm follows (with minor modifications) from the analysis in [18]; we skip the details.

Remarks: **(i)** A natural decomposition of the joint eMBB+URLLC scheduling is now apparent. On one hand, the eMBB scheduler maximizes utilities based on the *expected* channel rates stemming from uniform random puncturing of minislots (accounted for through the $(1-\rho)$ multiplicative factor), and does so using the iterative gradient scheduler. The URLLC scheduler, on the other-hand, is completely agnostic to either the channel state or the actual eMBB allocations and simply punctures minislots based on the current instantaneous demand.

(ii) The fact that the URLLC traffic placement is completely agnostic to the channel state and eMBB utilities/allocation is surprising. Intuitively it seems plausible that one could puncture an eMBB user with a lower marginal utility with more URLLC traffic, while protecting an eMBB user with a higher marginal utility and achieve a better sum utility. Further, it seems reasonable that eMBB users with a worse channel state (and thus lower rate) could be loaded with additional URLLC traffic. However, Theorem. 1 implies that there exists an optimal solution that is achieved by channel and utility oblivious and uniform random URLLC placement, thus providing a very simple algorithm for URLLC scheduling.

(iii) We remark that the optimality of random puncturing for linear loss models depends critically on the use of an opportunistic scheduler for eMBB traffic.To see this, consider a simple system with two symmetric eMBB users each with two possible channel states. The associated channel rates are either $\{2, 4\}$ packets/slot with equal probability, and independent across users and time slots. Suppose that we use a static (non-opportunistic) scheduler, which equally splits channel access between the users. It is easy to calculate that the rate to each user is then 1.5 packets/slot. Next suppose that the URLLC load is 50%, and that this traffic *randomly punctures* eMBB users. Then from symmetry, it follows that the rate per eMBB user is 0.75 packets/slot. In contrast, suppose that puncturing is opportunistic, where the user with the currently lower rate is punctured whenever possible (opportunistic puncturing of the currently worse eMBB user), a straightforward calculation shows that the rate to each eMBB user is 0.875 packets/slot, which is a *strict improvement over random puncturing.* At a high-level, this follows because opportunistic eMBB scheduling operates on the Pareto frontier of two-user capacity region, and consequently there is no residual opportunistic to be obtained by puncturing. However, with non-opportunistic scheduling, the system is not pushed to the boundary; thus, opportunistic puncturing can extract additional throughput for eMBB users.

IV. Convex Model – Minislot-Homogenous Policies

In this section we shall consider joint scheduling for wireless systems for convex superposition/puncturing loss models. This is a somewhat complex problem, whence we will focus our attention on a restricted, but still rich, class of scheduling policies which we refer to as minislot-homogeneous eMBB/URLLC schedulers. We identify a key concavity requirement in Assumption 2 (that is satisfied by convex loss functions) that enables a stochastic approximation approach for utility maximizing scheduling.

A. Minislot-homogeneous eMBB/URLLC Scheduling policies

We shall define minislot-homogeneous eMBB/URLLC schedulers as follows. First, feasible eMBB allocations $\boldsymbol{\phi} \in \Sigma$ will be restricted such that for any eMBB slot in channel state $s \in \mathcal{S}$ allocations are *minislot-homogeneous* across minislots in an eMBB slot, i.e., $\phi_{u,1}^s = \phi_{u,m}^s, \forall m \in \mathcal{M}$ and its overall allocation for the slot is given by $\phi_u^s = |\mathcal{M}|\phi_{u,1}^s$. The set of minislot-homogeneous eMBB allocations is thus given by

$$\Sigma^H := \left\{ \boldsymbol{\phi} \in \Sigma \mid u \in \mathcal{U},\ \phi_{u,m}^s = \phi_{u,1}^s \quad \forall m \in \mathcal{M}, \forall s \in \mathcal{S} \right\}.$$

Second, URLLC demand placements per minislot are done proportionally based on pre-specified weights, and these weights are assumed to be time-homogeneous across minislots. In particular such policies are parametrized by a weight matrix

$\boldsymbol{\gamma} \in \Sigma^H$, where the induced load on user u under channel state s and slot m is given by

$$L^s_{u,m} = \frac{\gamma^s_{u,m}}{\sum_{u' \in \mathcal{U}} \gamma^s_{u',m}} D(m) = \frac{\gamma^s_{u,1}}{f} D(m).$$

We shall call $\gamma^s_{u,1}$ the *URLLC placement factor* for eMBB user u in state s. The eMBB and URLLC allocations are coupled together since it must be the case that for all $u \in \mathcal{U}$ $L^s_{u,m} \leq \phi^s_{u,m} = \phi^s_{u,1}$ almost surely, i.e., one can not induce more superposition/puncturing load on a user than the resources it has been allocated on that slot. So the following condition must be satisfied. For all $m \in \mathcal{M}$ we have that

$$D(m) \leq \min_{u \in \mathcal{U}} \frac{\phi^s_{u,1}}{\gamma^s_{u,1}} f, \text{ almost surely.}$$

Recall that f denotes the maximum URLLC load per minislot so $D(m) \leq f$ almost surely, thus if $\frac{\phi^s_{u,1}}{\gamma^s_{u,1}} \geq 1$ the above condition will always hold. Yet if $\phi^s_{u,1} \geq \gamma^s_{u,1}$ for all u, then we have that $\phi^s_{u,1} = \gamma^s_{u,1}$, i.e., there is not flexibility to exploit careful placement of URLLC demands. Hence, we introduce the following assumption:

Assumption 1. *We say the system has a $(1-\delta)$ URLLC sharing factor per minislot if $D(m) \leq f(1-\delta)$ almost surely for all $m \in \mathcal{M}$, where $\delta \in (0,1)$.*

For any δ the above assumption implies that the *peak URLLC demand* in an eMBB slot can be at most $1-\delta$ which is lower than maximum possible value of one. Such an assumption is reasonable as we consider shared resources which are engineered to meet the peak URLLC loads while also serving eMBB traffic. Under a $(1-\delta)$ URLLC sharing factor a minislot-homogeneous eMBB resource allocation $\boldsymbol{\phi}$ and URLLC allocation $\boldsymbol{\gamma}$ is will be feasible if for all $s \in \mathcal{S}$ we have

$$(1-\delta) \leq \min_{u \in \mathcal{U}} \frac{\phi^s_{u,1}}{\gamma^s_{u,1}},$$

which is satisfied as long as $(1-\delta)\gamma^s_{u,1} \leq \phi^s_{u,1}$ for all $u \in \mathcal{U}$. This motivates the following definition:

Definition 1. *For a system with a $(1-\delta)$ sharing factor, the feasible minislot-homogeneous eMBB/URLLC scheduling policies are parameterized by $\boldsymbol{\phi}, \boldsymbol{\gamma} \in \Sigma^H$ such that $(1-\delta)\boldsymbol{\gamma} \leq \boldsymbol{\phi}$. We shall denote the set of such policies as follows:*

$$\Pi^{H,\delta} := \{(\boldsymbol{\phi}, \boldsymbol{\gamma}) \mid \boldsymbol{\phi}, \boldsymbol{\gamma} \in \Sigma^H \text{ and } (1-\delta)\boldsymbol{\gamma} \leq \boldsymbol{\phi}\},$$

where $\Pi^{H,\delta}$ is a convex set.

B. Characterization of the throughput region

In this section we characterize the throughput regions achievable under time-homogeneous scheduling.

Theorem 2. *For a system with a $(1-\delta)$ sharing factor and minislot-homogeneous scheduler $\boldsymbol{\pi} = (\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}) \in \Pi^{H,\delta}$ the average induced throughput for user $u \in \mathcal{U}$ in channel state $s \in \mathcal{S}$ is given by*

$$r^{\boldsymbol{\pi},s}_u = \mathrm{E}[f^s_u(\phi^{\boldsymbol{\pi},s}_u, \gamma^{\boldsymbol{\pi},s}_u D)],$$

and the overall average user throughputs are given by $\mathbf{c}^{\boldsymbol{\pi}} = (c^{\boldsymbol{\pi}}_u \mid u \in \mathcal{U})$ where $c^{\boldsymbol{\pi}}_u = \sum_{s \in \mathcal{S}} r^{\boldsymbol{\pi},s}_u p_S(s)$.

The proof is included in Appendix B. Based on the above we can define feasible throughput region constrained to the time-homogeneous policies in $\Pi^{H,\delta}$. First let us define

$$\mathcal{C}^{H,\delta} \quad = \quad \{\mathbf{c} \in \mathbb{R}^{|\mathcal{U}|}_+ \mid \exists \boldsymbol{\pi} \in \Pi^{H,\delta} \text{ s.t. } \mathbf{c} \leq \mathbf{c}^{\boldsymbol{\pi}}\}$$

and let $\hat{\mathcal{C}}^{H,\delta}$ denote the convex hull of $\mathcal{C}^{H,\delta}$. Note that rates in the convex hull are achievable through policies that do time sharing/randomization amongst minislot-homogeneous scheduling policies in $\Pi^{H,\delta}$.

Assumption 2. *For all $s \in \mathcal{S}$ and $u \in \mathcal{U}$ the functions $g^s_u(,)$ given by*

$$g^s_u(\phi^s_u, \gamma^s_u) = \mathrm{E}[f^s_u(\phi^s_u, \gamma^s_u D)], \tag{7}$$

are jointly concave on $\Pi^{H,\delta}$.

Lemma 1. *Assumption 2 is satisfied for systems where superposition/puncturing of each user is modelled via either a*

1) *Convex loss function or*
2) *Threshold loss function with fixed relative thresholds, i.e., $t^s_u(\phi^s_u) = \alpha^s_u$ for $\phi \in [0,1]$ and the URLLC demand distribution $F_D(\cdot)$ is such that $F_D(\frac{1}{x})$ is concave in x (satisfied by the truncated Pareto distribution).*

The proof is included in Appendix C. With this condition in place, we now describe the throughput region.

Theorem 3. *Under Assumption 2 we have that $\mathcal{C}^{H,\delta} = \hat{\mathcal{C}}^{H,\delta}$.*

The proof is available in the Appendix D. The above theorem implies that we do not have to consider time-sharing/randomization amongst minislot-homogeneous joint scheduling policies. Thus, with minislot-homogeneous policies and under the concavity of $g^{\pi,s}_u(\cdot,\cdot)$ from Assumption 2, the above result sets up a convex optimization problem in $(\boldsymbol{\phi}, \boldsymbol{\gamma})$, i..e, we have a concave cost function with convex constraints. Thus, by iteratively updating $(\boldsymbol{\phi}, \boldsymbol{\gamma})$, we can develop an online scheduling algorithm that asymptotically maximizes eMBB users' utility. This is descried next.

C. Stochastic approximation based online algorithm

We first restate the utility maximization problem for minislot-homogeneous URLLC/eMBB scheduling policies:

$$\max_{\boldsymbol{\phi}, \boldsymbol{\gamma} \in \Pi^{U,\delta}} \quad \sum_{u \in \mathcal{U}} U_u \left(\sum_{s \in \mathcal{S}} p_{\mathcal{S}}(s) g^s_u(\phi^s_u, \gamma^s_u) \right). \tag{8}$$

Observe that the objective function is concave because it consists of a sum of compositions of non-decreasing concave functions ($U_u(\cdot)$), and concave functions ($g^s_u(\cdot,\cdot)$) in $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}$ (if Assumption 2 holds). Further, the constraint set is convex. Therefore, the above problem fits in the framework of standard convex optimization problems. However, solving the above problem requires knowledge of all possible network states and their probability distribution, resulting in an *offline* optimization problem. In this section, we develop a stochastic

approximation based online algorithm to solve the above problem.

Online algorithm: Let $\overline{\boldsymbol{R}}(t-1) := \left(\overline{R}_1(t-1), \overline{R}_2(t-1), \ldots, \overline{R}_u(t-1), \ldots, \overline{R}_{|\mathcal{U}|}(t-1)\right)$ be the random vector denoting the average rates received by eMBB users up to eMBB slot $t-1$ under our online algorithm. Let $\overline{\boldsymbol{r}}(t-1)$ denote a realization of $\overline{\boldsymbol{R}}(t-1)$. Let s be the network state in slot t. Define vectors $\boldsymbol{\phi}^s := (\phi_u^s, \mid u \in \mathcal{U})$ and $\boldsymbol{\gamma}^s := (\gamma_u^s \mid u \in \mathcal{U})$. At the beginning of eMBB slot t, we compute vectors $\left(\tilde{\boldsymbol{\phi}}(t), \tilde{\boldsymbol{\gamma}}(t)\right)$ as the solution to the following optimization problem:

$$\max_{\phi^s, \gamma^s} \quad \sum_{u \in \mathcal{U}} U_u^{'}\left(\overline{r}_u(t-1)\right) g_u^s(\phi_u^s, \gamma_u^s), \tag{9}$$

$$\text{s.t.} \quad \boldsymbol{\phi}^s \geq (1-\delta)\,\boldsymbol{\gamma}^s, \tag{10}$$

$$\sum_{u \in \mathcal{U}} \phi_u^s = 1 \text{ and } \sum_{u \in \mathcal{U}} \gamma_u^s = 1, \tag{11}$$

$$\boldsymbol{\phi}^s \in [0,1]^{|\mathcal{U}|} \text{ and } \boldsymbol{\gamma}^s \in [0,1]^{|\mathcal{U}|}. \tag{12}$$

This optimization problem is a convex optimization problem and can be solved numerically using standard convex optimization techniques. Using $\left(\tilde{\boldsymbol{\phi}}(t), \tilde{\boldsymbol{\gamma}}(t)\right)$, we schedule URLLC and eMBB traffic as follows:

The eMBB scheduler: For notational ease, we fluidize the bandwidth. Specifically, we assume that the bandwidth of a resource block is very small when compared to the total bandwidth available. Hence, the bandwidth can be split into arbitrary fractions and we allocate fraction $\tilde{\phi}_u(t)$ of the total bandwidth to eMBB user u.

The URLLC Scheduler: We load different eMBB users with URLLC traffic according to the vector $\tilde{\boldsymbol{\gamma}}(t)$.

At the end of eMBB slot t, the eMBB scheduler receives feedback from the eMBB receivers indicating the rates received by the eMBB users. Let us denote the rate received eMBB user u in the slot by the random variable $R_u(t)$. We update $\overline{R}_u(t)$ as follows:

$$\overline{R}_u(t) = (1-\epsilon_t)\,\overline{R}_u(t-1) + \epsilon_t R_u(t), \tag{13}$$

where $\{\epsilon_t \mid t = 1, 2, 3, \ldots\}$ is a sequence of positive numbers which satisfy the following (standard) assumption:

Assumption 3. *The averaging sequence $\{\epsilon_t\}$ satisfies:*

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \quad \textit{and} \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty.$$

Finally, we state the main result of this section, which is the optimality of the stochastic approximation based online algorithm.

Theorem 4. *Let $\boldsymbol{r}^*$ be the optimal average rate vector received by eMBB users under the solution to the offline optimization problem. Suppose that Assumptions 3 and 2 hold, then we have that:*

$$\lim_{t \to \infty} \overline{\boldsymbol{R}}(t) = \boldsymbol{r}^* \quad \textit{almost surely}. \tag{14}$$

The proof is available in the Appendix E.

D. Optimality of Minislot-Homogeneous Policies

In the previous section we restricted ourselves to minislot-homogeneous policies. In this section will justify this choice. Let us consider a generalization of minislot-homogeneous policies where the URLLC placement in each minislot can depend on the history of URLLC arrivals prior to that minislot. Such a policy will obviously perform better than minislot-homogeneous URLLC placement policies since in a minislot-homogeneous policy we decide the URLLC placement at the beginning of an eMBB slot based on the expected loss due to puncturing/superposition and do not adapt it based on the realization of URLLC demands per minislot. However, finding an optimal scheduling policy under this generalization can be computationally expensive as compared to minislot-homogeneous policies which are attractive due to their simplicity. In this section we identify conditions under which minislot-homogeneous URLLC placement polices perform as well as the general class of *causal* and *minislot-dependent* policies. These terms are defined below.

Definition 2. A scheduler is said to be *causal* if at the beginning of a mini-slot m the scheduler knows the realizations of $D(1), D(2), \ldots, D(m-1)$ and is unaware of the realizations of $D(m), D(m+1), \ldots, D(|\mathcal{M}|)$.

Definition 3. A scheduling policy is said to be *minislot-dependent* if the URLLC placement policy can vary with the minislot index m and previous URLLC demands in the eMBB slot.

The decision variables in a causal and minislot-dependent joint scheduling policy π can be described as follows:

1) At the beginning of an eMBB slot, the scheduler chooses $\phi_u^{\pi,s}, u \in \mathcal{U}$ such that
$$\sum_{u \in \mathcal{U}} \phi_u^{\pi,s} = 1 \text{ and } \phi_u^{\pi,s} \in [0,1] \quad \forall u \in \mathcal{U}. \tag{15}$$
2) In each mini-slot m, the total puncturing placed on eMBB user u is given by $\gamma_{u,m}^{\pi,s}\left(\boldsymbol{d}^{(1:m-1)}\right) D_m$, where $\gamma_{u,m}^{\pi,s}(\cdot)$ characterizes the URLLC placement in minislot m as function of the previously seen URLLC demands $\boldsymbol{D}^{(1:m-1)} := (D(1), D(2), \ldots, D(m-1))$. Let $\boldsymbol{d}^{(1:m-1)}$ is a realization of $\boldsymbol{D}^{(1:m-1)}$. For any m and $\boldsymbol{d}^{(1:m-1)}$, $\gamma_{u,m}^{\pi,s}\left(\boldsymbol{d}^{(1:m-1)}\right)$ has to satisfy the following constraints.
$$\sum_{u \in \mathcal{U}} \gamma_{u,m}^{s,\pi}(\boldsymbol{d}^{(1:m-1)}) = 1, \tag{16}$$
$$\gamma_{u,m}^{s,\pi}(\boldsymbol{d}^{(1:m-1)}) \leq \frac{\phi_u^{\pi,s}}{|\mathcal{M}|\,(1-\delta)} \quad \forall u \in \mathcal{U}, \tag{17}$$
$$\gamma_{u,m}^{s,\pi}(\boldsymbol{d}^{(1:m-1)}) \in [0,1] \quad \forall u \in \mathcal{U}. \tag{18}$$

Observe that the URLLC placement factor for causal and minislot-dependent scheduling policy is not just dependent on the user and network state but it also depends on the mini-slot index and past URLLC demands.

Let $\tilde{\Pi}$ be the set of all causal and mini-slot dependent scheduling policies. In our online algorithm (9), for any

eMBB slot t, we find the policy which solves the following optimization problem with non-negative weights w_u.

$$\mathcal{OP}_1: \quad \max_{\pi\in\tilde{\Pi}}: \sum_{u\in\mathcal{U}} w_u g_u^{\pi,s}\left(\phi_u^{\pi,s}, \boldsymbol{\gamma}_u^{\pi,s}\right), \tag{19}$$

where s is the current network state, $\boldsymbol{\gamma}_u^{\pi,s} := \left(\gamma_{u,1}^{\pi,s}(\cdot), \gamma_{u,2}^{\pi,s}(\cdot), \ldots, \gamma_{u,|\mathcal{M}|}^{\pi,s}(\cdot)\right)$ is the vector of URLLC placement factors of all minislots (with slight abuse of notation) and $g_u^{\pi,s}(\cdot,\cdot)$ is the average rate experienced by eMBB user u under policy π. $g_u^{\pi,s}(\cdot,\cdot)$ is given by the following expression:

$$g_u^{\pi,s}\left(\phi_u^{\pi,s}, \boldsymbol{\gamma}_u^{\pi,s}\right) := r_u^s \phi_u^{\pi,s} \mathbb{E}\left[1 - h_u^s\left(\frac{\sum_{m=1}^{|\mathcal{M}|}\gamma_{u,m}^{\pi,s}\left(\boldsymbol{D}^{(1:m-1)}\right)D_m}{\phi_u^{\pi,s}}\right)\right], \tag{20}$$

where the expectation is computed with respect to the joint distribution of $D(1)$, $D(2)$,..., $D(|\mathcal{M}|)$. One can formulate the above optimization problem as a Markov Decision Problem (MDP), however the state space for such an MDP is prohibitively large. Furthermore we note that minislot-homogeneous policies are attractive in terms of it computational complexity. In general, one cannot expect optimal minislot-homogeneous policies to perform as well as optimal minislot dependent policies, however, if we restrict ourselves to convex *homogeneous* loss functions, then we can show that minislot-homogeneous policies are in fact optimal over $\tilde{\Pi}$.

Definition 4. A loss function $h_u^s(\cdot)$ is said to be homogeneous if there exists a real number p such that $\forall\, x \in [0,1]$ and $\kappa \geq 0$ we have that

$$h_u^s(\kappa x) = \kappa^p h_u^s(x). \tag{21}$$

Even with this restriction we can model useful loss functions which could possibly be user and network state dependent. Some examples are given below.

1) **Linear:** $h_u^s(x) = k_u^s(x)$, where $k_u^s \geq 0$.
2) **Monomial:** $h_u^s(x) = k_u^s(x)^q$ where $k_u^s \geq 0$ and $q \geq 1$.

Our main result on the optimality of minislot-homogeneous policies is proved in Appendix F and stated next.

Theorem 5. *If the support of URLLC demands D is a finite discrete set and eMBB loss functions are homogeneous and convex, then there exists an optimal solution $(\boldsymbol{\phi}^{s,*}, \boldsymbol{\gamma}^{s,*}(\cdot))$ for $\mathcal{OP}_1$ with a minislot-homogeneous URLLC placement policy $\boldsymbol{\gamma}^{s,*}$.*

E. Optimal eMBB Slot Slicing

In Section II we have used uniform slot sizes for eMBB users, i.e. the allocated minislots to all users span the entire width of the slot (see Figure 4; henceforth referred to as frequency slices). However, new proposals allow greater flexibility in slot allocation, e.g., the capability to choose different slices over both time and frequency for different eMBB users [1]. In this section we will show that while it is possible to slice eMBB users' resources flexibly, it is preferable to slice frequency (see 4) than time from the point of view of puncturing losses for convex loss functions.

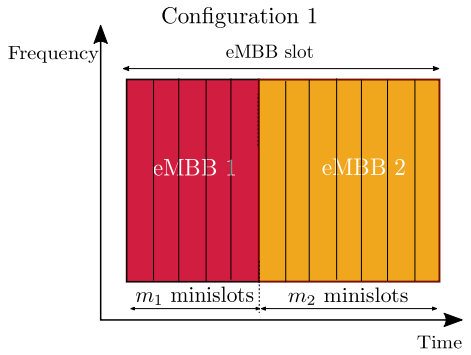

Fig. 3. Time Slices: In this configuration, eMBB users share resources over time in an eMBB slot.

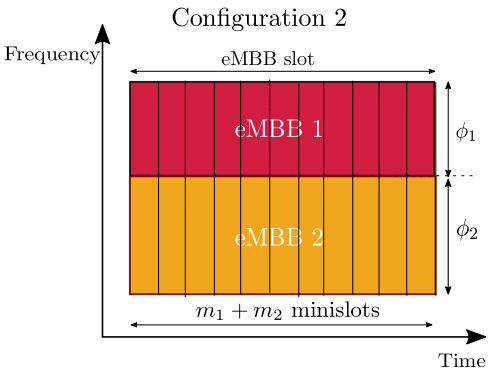

Fig. 4. Frequency Slices: In this configuration, eMBB users homogeneously share frequency in an eMBB slot.

The essence of the discussion can be captured by comparing the two resource allocation configurations shown in Figures 3 and 4. In Configuration 1 (time slices), eMBB user 1 is allocated the entire frequency band for a subset of m_1 minislots. Similarly eMBB user 2 is allocated the entire frequency band for its subset of m_2 minislots. The network state s is assumed to be the same for the entire m_1+m_2 minislots. This implies that the loss functions of eMBB users $(h_u^s(\cdot))$ do not change throughout the m_1+m_2 minislots. In Configuration 2 (frequency slices) we allocate an eMBB user 1 a fraction ϕ_1 of the bandwidth for a duration of m_1+m_2 minislots, where $\phi_1 := \frac{m_1}{m_1+m_2}$ and similarly for eMBB user 2. Note that the total resources allocated to eMBB users, which is represented by the area allocated in the time-frequency plane is same in both configurations.

In Configuration 1, the total puncturing observed by eMBB user 1 is given by $\sum_{m=1}^{m_1} D(m)$ and similarly for eMBB user 2. Whereas in Configuration 2, under uniform URLLC placement, the total puncturing observed by eMBB user 1 is given by $\sum_{m=1}^{m_1+m_2}\phi_1 D(m)$. Note that the mean total puncturing is same in both the configurations.

The main result of this section is given below:

Theorem 6. *Under the assumption of i.i.d. URLLC demands*[5] *($D(m)$, $m = 1, 2, \ldots, m_1 + m_2$) and convex loss functions ($h_u^s(\cdot)$), for any eMBB user, e.g., eMBB user* 1*, we have that*

$$\mathbb{E}\left[h_1^s\left(\sum_{m=1}^{m_1} D(m)\right)\right] \geq \mathbb{E}\left[h_1^s\left(\sum_{m=1}^{m_1+m_2} \phi_1 D(m)\right)\right]. \quad (22)$$

Proof of this result is given in Appendix G.

Remarks: The above theorem shows that the expected loss suffered by an eMBB user due to URLLC puncturing in Configuration 1 (time slicing) is higher than in Configuration 2 (frequency slicing). This implies that it is preferable for eMBB users to spread their resource allocation over time from the perspective of reducing their loss due to puncturing. The underlying reason is that Configuration 2 results in smaller variability in the total puncturing even though both the configurations have the same mean total puncturing. Since the loss functions are convex, a lower variability leads to a lower expected loss. Finally, for more complex (rectangular) slices, we can now apply Thm. 6 iteratively and show that using frequency slices with appropriate scaling of the bandwidth allocation results in a higher average rate for eMBB users.

V. Threshold Model and Placement Policies

In the previous section, we developed a stochastic approximation based algorithm for minislot-homogeneous policies. This algorithm iteratively solves the optimization problem given in (9). This optimization problem jointly optimizes over a pair of row vectors $(\boldsymbol{\phi}^s, \boldsymbol{\gamma}^s)$. While this convex optimization problem can be solved using standard methods, it could become computationally challenging as the number of users increases.

In this section, we shall restrict our attention to a threshold model for superposition/puncturing, and look at policies that impose structural conditions on the puncturing matrix $\boldsymbol{\gamma}$. We will show that the resulting class of policies have nice theoretical properties that lead to simpler online algorithms (solving (4), which is an one-dimensional search).

We consider two types of structural conditions on $\boldsymbol{\gamma}$:

(i) Resource Proportional (RP) Placement: The first is based on allocating URLLC demands in proportion to eMBB user slot allocations, i.e., $\gamma_u^s = \phi_u^s$. We refer to this as Resource Proportional (RP) Placement and denote such policies by

$$\Pi^{RP,\delta} := \{(\boldsymbol{\phi}, \boldsymbol{\gamma}) \in \Pi^{H,\delta} \mid \boldsymbol{\gamma} = \boldsymbol{\phi}\},$$

and define the associated achievable throughput region

$$\mathcal{C}^{RP,\delta} \quad = \quad \{\mathbf{c} \in \mathbb{R}_+^{|\mathcal{U}|} \mid \exists \boldsymbol{\pi} \in \Pi^{RP,\delta} \text{ s.t. } \mathbf{c} \leq \mathbf{c}^{\boldsymbol{\pi}}\}.$$

The motivation for RP Placement comes from the optimality of random placement for the linear model in Section III. Observe that if puncturing occurs uniformly randomly, then the expected number of punctures is directly proportional to the fraction of bandwidth allocated to an eMBB user. Thus, RP Placement can be viewed as a *determinized version* of the random placement strategy which ensures that the proportions of puncturing satisfy resource proportional ratios.

(ii) Threshold Proportional (TP) Placement: The second policy allocates URLLC demands in proportion to the eMBB users associated loss thresholds so as to avoid losses,

$$\gamma_u^s = \frac{\phi_u^s t_u^s(\phi_u^s)}{\sum_{u' \in \mathcal{U}} \phi_{u'}^s t_{u'}^s(\phi_{u'}^s)}.$$

We refer to this as Threshold Proportional (TP) Placement and denote such policies by

$$\Pi^{TP,\delta} := \{(\boldsymbol{\phi}, \boldsymbol{\gamma}) \in \Pi^{H,\delta} \mid \gamma_u^s = \frac{\phi_u^s t_u^s(\phi_u^s)}{\sum_{u' \in \mathcal{U}} \phi_{u'}^s t_{u'}^s(\phi_{u'}^s)} \forall s \in \mathcal{S}, u \in \mathcal{U}\}.$$

The associated achievable throughput region is denoted

$$\mathcal{C}^{TP,\delta} \quad = \quad \{\mathbf{c} \in \mathbb{R}_+^{|\mathcal{U}|} \mid \exists \boldsymbol{\pi} \in \Pi^{TP,\delta} \text{ s.t. } \mathbf{c} \leq \mathbf{c}^{\boldsymbol{\pi}}\}.$$

First we state a corollary to Theorem 2 which characterizes the rates under different URLLC placement policies for systems having threshold loss model for superposition/puncturing.

Corollary 1. *Under a $(1-\delta)$ sharing factor and time-homogeneous scheduler $\boldsymbol{\pi} = (\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}) \in \Pi^{H,\delta}$ the probability of induced eMBB loss for user $u \in \mathcal{U}$ in channel state $s \in \mathcal{S}$ is given by*

$$\epsilon_u^{\boldsymbol{\pi},s} = 1 - F_D\left(\frac{\phi_u^{\boldsymbol{\pi},s} t_u^s(\phi_u^{\boldsymbol{\pi},s})}{\gamma_u^{\boldsymbol{\pi},s}}\right).$$

where F_D denotes the cumulative distribution function of the URLLC demands on a typical eMBB slot. Then the associated user throughput is given by

$$r_u^{\boldsymbol{\pi},s} = \hat{r}_u^s \phi_u^{\boldsymbol{\pi},s} F_D\left(\frac{\phi_u^{\boldsymbol{\pi},s} t_u^s(\phi_u^{\boldsymbol{\pi},s})}{\gamma_u^{\boldsymbol{\pi},s}}\right).$$

and the overall user throughputs are given by $\mathbf{c}^{\boldsymbol{\pi}} = (c_u^{\boldsymbol{\pi}} : u \in \mathcal{U})$ where

$$c_u^{\boldsymbol{\pi}} = \sum_{u \in \mathcal{U}} \hat{r}_u^s \phi_u^s F_D\left(\frac{\phi_u^{\boldsymbol{\pi},s} t_u^s(\phi_u^{\boldsymbol{\pi},s})}{\gamma_u^{\boldsymbol{\pi},s}}\right) p_S(s).$$

The following two corollaries are direct consequences of Corollary 1 and Theorem 3 restricted to RP and TP Placement strategies, and characterize the capacity regions under the two policies.

Corollary 2. *Consider a wireless system with full sharing factor and time-homogeneous scheduler based on the RP URLLC Placement policy $\boldsymbol{\pi} = (\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}) \in \Pi^{RP,\delta}$. Then any eMBB resource allocation $\boldsymbol{\phi}$ combined with a RP URLLC demand placement policy, $\boldsymbol{\gamma} = \boldsymbol{\phi}$ is feasible. The probability of loss for user $u \in \mathcal{U}$ in channel state $s \in \mathcal{S}$ is given by*

$$\epsilon_u^{\boldsymbol{\pi},s} = 1 - F_D(t_u^s(\phi_u^{\boldsymbol{\pi},s})),$$

with associated user throughput

$$r_u^{\boldsymbol{\pi},s} \quad = \quad \hat{r}_u^s \phi_u^s F_D(t_u^s(\phi_u^{\boldsymbol{\pi},s})). \quad (23)$$

[5]This result can be extended to exchangeable URLLC demands. We use i.i.d. assumption to maintain consistency with other sections.

Further if for all $s \in \mathcal{S}$ and $u \in \mathcal{U}$ the functions $g_u^s(,)$ given by

$$g_u^s(\phi_u^s) \quad = \quad \phi_u^s F_D(t_u^s(\phi_u^{\boldsymbol{\pi},s})), \tag{24}$$

are concave then $\mathcal{C}^{RP,\delta} = \hat{\mathcal{C}}^{RP,\delta}$.

Corollary 3. *Under a $(1-\delta)$ sharing factor and jointly uniform scheduler based on the TP URLLC Placement policy $\boldsymbol{\pi} = (\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}) \in \Pi^{TP,\delta}$, the probability of induced eMBB loss user $u \in \mathcal{U}$ in channel state $s \in \mathcal{S}$ is given by*

$$\epsilon_u^{\boldsymbol{\pi},s} \quad = \quad 1 - F_D(\sum_{u \in \mathcal{U}} \phi_u^{\boldsymbol{\pi},s} t_u^s(\phi_u^{\boldsymbol{\pi},s})), \tag{25}$$

with associated user throughput

$$r_u^{\boldsymbol{\pi},s} = \hat{r}_u^s \phi_u^s F_D(\sum_{u \in \mathcal{U}} \phi_u^{\boldsymbol{\pi},s} t_u^s(\phi_u^{\boldsymbol{\pi},s})). \tag{26}$$

Further if for all $s \in \mathcal{S}$ and $u \in \mathcal{U}$ the functions $g_u^s(,)$ given by

$$g_u^s(\phi_u^s, \gamma_u^s) \quad = \quad \phi_u^s F_D(\sum_{u \in \mathcal{U}} \phi_u^{\boldsymbol{\pi},s} t_u^s(\phi_u^{\boldsymbol{\pi},s})), \tag{27}$$

are jointly concave then $\mathcal{C}^{TP,\delta} = \hat{\mathcal{C}}^{TP,\delta}$.

The following theorem provides a formal motivation for TP Placement. The main takeaway here is that the *probability of any loss in an eMBB slot under TP Placement policy is a lower bound for all other strategies.* Note that minimizing the probability of any eMBB loss is not same as minimizing eMBB rate loss.

Theorem 7. *Consider a system with $(1-\delta)$ sharing factor. Consider a joint scheduling policy based on the TP URLLC placement i.e, $\boldsymbol{\pi} = (\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}) \in \Pi^{TP,\delta}$. Then $\boldsymbol{\pi}$ achieves the minimum probability of any eMBB loss amongst all joint scheduling policies using the same eMBB resource allocation $\boldsymbol{\phi}^{\boldsymbol{\pi}}$.*

The proof is included in Appendix H.

Next we consider online algorithms that implement the RP and TP Placement policies. While the stochastic approximation algorithm developed in Section IV-C can clearly be used, the additional structure imposed by the RP and TP Placement policies, and the shape of the threshold loss function (discussed below) can result in much simpler algorithms (with optimality guarantees).

We consider the case where $t_u^s(\phi)$ is a (state dependent but ϕ independent) constant, i.e., $t_u^s(\phi) = \alpha^s$, where $\alpha^s \in (0,1)$. Intuitively, this means that eMBB traffic which has a higher share of the bandwidth is more resilient to losses (e.g. through coding over larger fraction of resources). Then, by substituting this loss function in (23) and (26) (where we also use the fact that $\sum_{u \in \mathcal{U}} \phi_u^s = 1$), we have that

$$r_u^{\boldsymbol{\pi},s} = \hat{r}_u^s \phi_u^s F_D(\alpha^s).$$

Comparing with the development in Section III-B, we observe that the cost and constraints are identical if $F_D(\alpha^s)$ replaces $(1-\rho)$. Note that a small difference is that $F_D(\alpha^s)$ is state dependent, whereas $(1-\rho)$ does not depend on the state; however, it is easy to see that the development in Section III-B immediately generalizes to this setting. Hence, we can interpret $F_D(\alpha^s)$ as the state dependent average rate loss due to puncturing via the RP or TP Placement policies.

We can now employ the rate-based iterative gradient scheduler developed in Section III-B (by replacing $(1-\rho)$ in (5) by a user-dependent $F_D(\alpha^s)$), and the theoretical guarantees directly carry over. As this algorithm only minimizes over users at each slot in (4), this is easier to implement when compared to the stochastic approximation algorithm developed in Section IV-C.

VI. Simulations

We consider a system with a total of 100 RBs available per eMBB slot, and with 8 minislots per eMBB slot. In an eMBB slot, $\hat{r}_u^s$ for an eMBB user is drawn from the finite set $\{1,2,3,4,5,6,7,8,9,10\}$ Mbps according to a probability distribution and i.i.d. across users and slots. Our system consists of 20 users, and with 100 channel states (all equally likely). The (20 users $\times$ 100 states) rate matrix is one-time synthesized by independently and uniformly sampling a rate from the finite rate set for each matrix element. For 10 eMBB users, we have chosen the probability distribution such that the average rate is 7 Mbps. For the rest, probability distribution is such that the average rate is 3 Mbps. This models two classes of users, one class with higher link rates which can tolerate a higher amount of puncturing and the other with lower link rates which can tolerate lesser amount of puncturing. This is reasonable as a user with a higher channel rate can code more robustly and protect its transmissions from URLLC puncturing more than a user with a lower channel rate. In this spirit we shall call users with 7 Mbps average rates as 'robust' users and users with 3 Mbps average rates as 'sensitive' users. We use the utility function $U_u(r) = \log(r)$ for all users.

We first show that joint scheduling is necessary to preserve eMBB throughputs. To that end we benchmark our optimal online algorithm (stochastic approximation algorithm, see Section IV-C) for convex loss functions with a scheme which performs standard gradient based scheduling for eMBB users and Resource Proportional (RP) URLLC placement. Note that for convex loss functions, RP placement strategy does not take into account the eMBB user's sensitivity to delays. For users with average rate 7 Mbps, we use the loss function $h_u^s(x) = x^2$. For users with average rate 3 Mpbs, we use the following loss function:

$$h_u^s(x) = \begin{cases} \left(\frac{x}{0.7}\right)^2, & \text{if } x \leq 0.7, \\ 0, & \text{if } 0.7 < x \leq 1. \end{cases} \tag{28}$$

URLLC demands in a minislot is drawn from a binomial distribution which can take values 0 with p and $\frac{1-\delta}{8}$ with probability $1-p$. Note that this ensures that peak URLLC load in an eMBB slot is less than or equal to $1-\delta$.

In Fig. 5, we compare the average sum utility under our optimal joint scheduler and the RP based policy as a function of the URLLC load. As the load increases, RP performs

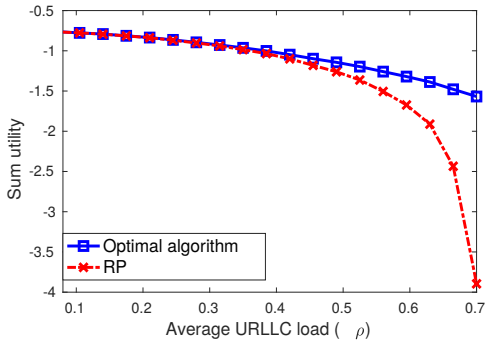

Fig. 5. Sum utility as a function of URLLC load ρ for the optimal and RP policies under convex model $\delta = 0.3$.

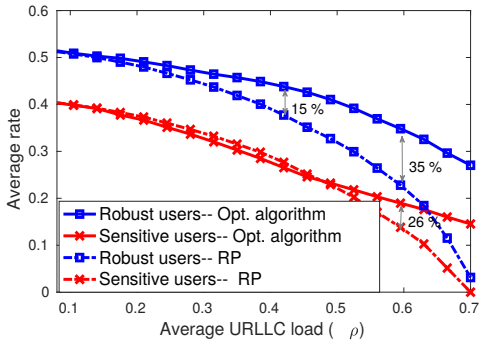

Fig. 6. Average rates as a function of URLLC load ρ for the optimal and RP policies under convex model $\delta = 0.3$.

poorly. To understand this phenomenon in detail, we have plotted the average rates of robust and sensitive users under the two policies in Fig. 6. As we increase the URLLC load, the average eMBB rates of both sensitive and robust users decrease rapidly. For example, when $\rho = 0.4$, RP has 15 % lower throughput for robust users and almost similar performance for sensitive users as compared to optimal algorithm. Further as we increase ρ to 0.6, the throughput of robust and sensitive users in RP decrease by 35 % and 26 %, respectively.

Sensitive users are the most affected by URLLC puncturing. When the RP URLLC placement policy is combined with the standard gradient based algorithm for eMBB users, it allocates more resources to sensitive users because they have higher marginal utility. Since sensitive users receive more bandwidth, under the RP URLLC placement strategy they receive more puncturing. This will lead to even more allocation of resources to sensitive users and this process continues until robust users have similar marginal utilities (due to reduced rates) as sensitive users. Hence, the robust users are resource starved. As we increase the URLLC load further, sensitive users receive even more URLLC puncturing and neither the robust nor sensitive users get good average rates when compared to the optimal joint scheduler. This shows that we require joint scheduling of eMBB and URLLC to exploit the heterogeneity in sensitivities to URLLC puncturing in maximizing eMBB utilities.

Next we consider a threshold based loss model with $\alpha^s = 0.3$ for 50% of eMBB states and $\alpha^s = 0.7$ for the rest. We use the utility function $U_u(r) = \log(r) + 6.5$ for all eMBB users, where r is measured in Mbps (constant added to ensure non-negativity of the sum utility). URLLC load in an eMBB slot (D) is generated based on the truncated Pareto distribution with tail exponent $\eta = 2$. We compare the optimal policy (stochastic approximation algorithm, see Section IV-C) with that from the TP Placement policy (the simpler gradient algorithm in Section V). In this case, since the threshold functions are (state-dependent) constants, the RP and TP Placement policies are the same. As we can see in Figure 7, unlike the convex loss model the RP/TP Placement policy tracks the optimal policy very well.

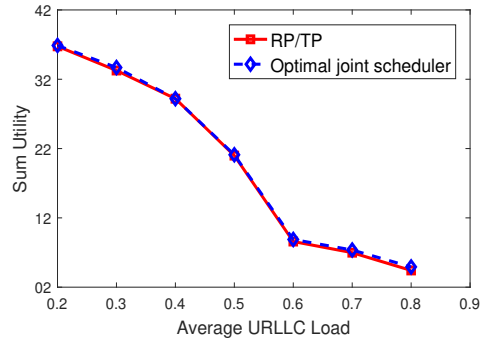

Fig. 7. Sum utility as a function of URLLC load ρ for the optimal and TP Placement policies under threshold model ($\delta = 0.1$).

In Figure 9, we study the trade-off between achieving a higher eMBB utility and lowering the mean delay of URLLC traffic for different values of the sharing factor $1-\delta$. Figure 9 plots the corresponding probability that the URLLC traffic delay exceeds two minislots ($0.125 \times 2 = 0.25$ msec). To study this trade-off we generate URLLC arrivals in each minislot from an uniform distribution between $[0, 1/8]$ (recall there are

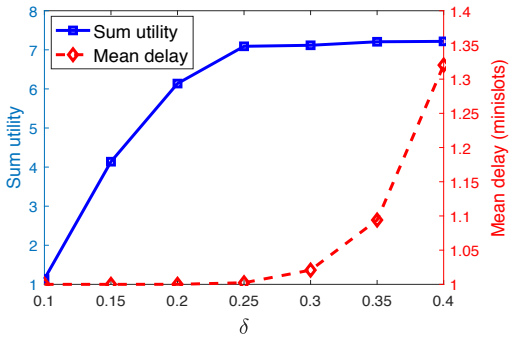

Fig. 8. Sum utility and mean URLLC delay as a function of δ.

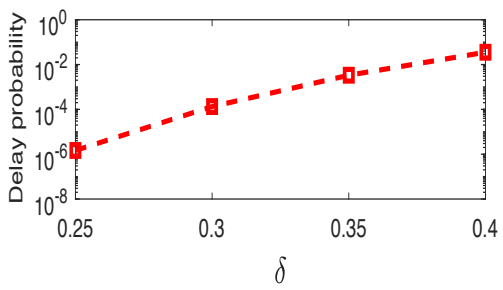

Fig. 9. Log-scale plot of the probability that URLLC traffic is delayed by more than two minislots (0.25 msec) for various values of δ.

8 minislots). In each minislot, we can serve at most $\frac{1-\delta}{8}$ units of URLLC traffic. If the URLLC load in a given minislot is more than $\frac{1-\delta}{8}$, the remaining URLLC traffic is queued and served in the next minislot on a FCFS basis. For the eMBB users we use a convex model with $h_u^s(x) = e^{\kappa_u(x-1)}$ where κ_u determines the sensitivity of an eMBB user to an URLLC load. We have chosen $\kappa = 0.2$ for 50 % of the users and $\kappa = 0.7$ for the rest. We also set $\forall u\ U_u(x) = \log(x) + 4.2$ (constant added to ensure positive sum utility). In summary, a larger value of δ limits the amount of URLLC traffic than can be served in a minislot. However, a larger δ enlarges the constraint set $\Pi^{H,\delta}$ in the eMBB utility maximization problem, and hence we get higher eMBB utility.

VII. Conclusion

In this paper, we have developed a framework and algorithms for joint scheduling of URLLC (low latency) and eMBB (broadband) traffic in emerging 5G systems. Our setting considers recent proposals where URLLC traffic is dynamically multiplexed through puncturing/superposition of eMBB traffic. Our results show that this joint problem has structural properties that enable clean decompositions, and corresponding algorithms with theoretical guarantees.

Acknowledgements

The work of Arjun Anand was partially supported by FutureWei Technologies and NSF grant CNS-1731658, Gustavo de Veciana was partially supported by NSF grants CNS-1343383 and CNS-1731658, and Sanjay Shakkottai was partially supported by NSF grants CNS-1343383 and CNS-1731658, and the US DoT D-STOP Tier 1 University Transportation Center.

References

[1] 3GPP TSG RAN WG1 Meeting 87, November 2016.
[2] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5g wireless networks," in *Proc. INFOCOM*, May 2018.
[3] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 53–59, March 2016.
[4] Chairman's notes 3GPP: 3GPP TSG RAN WG1 Meeting 88bis, Available at http://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_88b/Report/, April 2017.
[5] R. Srikant and L. Ying, *Communication Networks: An Optimization, Control, and Stochastic Networks Perspective*. Cambridge University Press, 2014.
[6] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, 2006.
[7] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S. A. Ashraf, J. Huschke, I. Aktas, and J. Ansari, "Wireless communication for factory automation: an opportunity for LTE and 5G systems," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 36–43, June 2016.
[8] O. N. C. Yilmaz, Y. P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, June 2015, pp. 1190–1195.
[9] M. Gidlund, T. Lennvall, and J. Akerberg, "Will 5G become yet another wireless technology for industrial automation?" in *2017 IEEE International Conference on Industrial Technology (ICIT)*, March 2017, pp. 1319–1324.
[10] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.
[11] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, Sept 2016.
[12] G. Durisi, T. Koch, J. Ostman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna rayleigh-fading channels," *IEEE Trans. on Comm.*, vol. 64, no. 2, pp. 618–629, Feb 2016.
[13] B. Singh, Z. Li, O. Tirkkonen, M. A. Uusitalo, and P. Mogensen, "Ultra-reliable communication in a factory environment for 5G wireless networks: Link level and deployment study," in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sept 2016, pp. 1–5.
[14] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource Optimization with Flexible Numerology and Frame Structure for Heterogeneous Services," *ArXiv e-prints*, 2018.
[15] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," *ArXiv e-prints*, 2018.
[16] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN Uplink: An Information-Theoretic Study," *ArXiv e-prints*, 2018.
[17] D. Julian, "Erasure networks," in *Proceedings IEEE International Symposium on Information Theory,*, Jul. 2002.
[18] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations Research*, vol. 53, no. 1, pp. 12–25, 2005.
[19] S. Bodas, S. Shakkottai, L. Ying, and R. Srikant, "Low-complexity scheduling algorithms for multi-channel downlink wireless networks," in *Proceedings of IEEE Infocom*, 2010.
[20] ——, "Scheduling for small delay in multi-rate multi-channel wireless networks," in *Proceedings of IEEE Infocom*, 2011.
[21] S. Boyd and L. Vandenberge, *Convex Optimization*. Cambridge University Press, 2003.
[22] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. Springer, 1997.

APPENDIX

A. Proof of Theorem 1

Clearly since $\Pi^{LR} \subset \Pi$ we have that $\mathcal{C}^{LR} \subset \mathcal{C}$

Now consider any policy $\pi \in \Pi$ with eMBB user allocations $\boldsymbol{\phi}^{\pi}$ and URLLC loads $\bar{\mathbf{l}}^{\pi}$ and associated long term throughput is $\mathbf{c}^{\pi}$ given by

$$c_u^{\pi} = \sum_{s \in \mathcal{S}} \hat{r}_u^s (\phi_u^{\pi,s} - \bar{l}_u^{\pi,s}) p_S(s).$$

Let us define a π' based on π to have per minislot eMBB user allocations given by

$$\phi_{u,m}^{\pi',s} = \frac{\phi_u^{\pi,s} - \bar{l}_u^{\pi,s}}{\sum_{u' \in \mathcal{U}} \phi_{u'}^{\pi,s} - l_{u'}^{\pi,s}} f = \frac{\phi_u^s - \bar{l}_u^{\pi,s}}{1-\rho} f,$$

for $s \in \mathcal{S}$, $u \in \mathcal{U}$ and $m \in \mathcal{M}$. Since induced mean loads on an eMBB user can not exceed its allocation we have that $\boldsymbol{\phi}^{\pi} \geq \bar{\mathbf{l}}^{\pi}$ so the above allocations are positive. Note also that this allocation is not minislot dependent, but normalized so that per minislot they sum to f and over the whole eMBB slot sum to 1, i.e., $\boldsymbol{\phi}^{\pi'} \in \Sigma$. Thus for such an allocation we have that

$$\phi_u^{\pi',s} = \frac{\phi_u^s - \bar{l}_u^{\pi,s}}{1-\rho}.$$

Also suppose that π' uses randomized URLLC placement across minislots which induces mean URLLC loads proportional to the allocations, i.e., $\bar{l}_u^{\pi',s} = \rho \phi_u^{\pi',s}$. It follows that

$$\begin{aligned} \phi_u^{\pi',s} - \bar{l}_u^{\pi',s} &= \phi_u^{\pi',s} - \rho \phi_u^{\pi',s} \\ &= (1-\rho)\phi_u^{\pi',s} \\ &= \phi_u^{\pi,s} - \bar{l}_u^{\pi,s}, \end{aligned}$$

and so $c_u^{\pi,s} = c_u^{\pi',s}$ for all $s \in \mathcal{S}$ and $u \in \mathcal{U}$. Thus for any policy π there is a policy π' which uses randomized URLLC placement and achieves the same long term throughputs. It follows that $\mathcal{C} \subset \mathcal{C}^{LR}$ and so $\mathcal{C} = \mathcal{C}^{LR}$.

B. Proof of Theorem 2

Under a policy $\boldsymbol{\pi} = (\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}) \in \Pi^{H,\delta}$ we have that the induced loads are given by

$$L_{u,m}^{\boldsymbol{\pi},s} = \frac{\gamma_{u,1}^{\boldsymbol{\pi},s}}{f} D(m),$$

so we have that

$$L_u^{\boldsymbol{\pi},s} = \sum_{u \in \mathcal{U}} L_{u,m}^{\boldsymbol{\pi},s} = \frac{\gamma_{u,1}^{\boldsymbol{\pi},s}}{f} \sum_{u \in \mathcal{U}} D(m) = \frac{\gamma_{u,1}^{\boldsymbol{\pi},s}}{f} D = \gamma_u^{\boldsymbol{\pi},s} D.$$

where the last equality follows from the uniformity of URLLC splits and normalization it follows that

$$r_u^{\boldsymbol{\pi},s} = \mathrm{E}[f_u^s(\phi_u^{\boldsymbol{\pi},s}, L_u^{\boldsymbol{\pi},s})] = \mathrm{E}[f_u^s(\phi_u^{\boldsymbol{\pi},s}, \gamma_u^{\boldsymbol{\pi},s} D)].$$

C. Proof of Lemma 1

Recall that convex loss functions are specified as follows

$$f_u^s(\phi_u^s, l_u^s) = \hat{r}_u^s \phi_u^s (1 - h_u^s \left(\frac{l_u^s}{\phi_u^s} \right)),$$

with $h_u^s : [0,1] \to [0,1]$ a convex increasing function. For time-homogenous policies we have defined

$$\begin{aligned} g_u^s(\phi_u^s, \gamma_u^s) &= \mathrm{E}[f_u^s(\phi_u^s, \gamma_u^s D)] \\ &= \hat{r}_u^s E[\phi_u^s - \phi_u^s h_u^s (\frac{\gamma_u^s}{\phi_u^s} D)]. \end{aligned}$$

Recall that convex function $h(\cdot)$ one can define a function $l(\phi, \gamma) = \phi h(\frac{\gamma}{\phi})$ known as the perspective of $h(\cdot)$ which is known to be jointly convex in its arguments. It follows that $\phi - \phi h(\frac{\gamma}{\phi})$ is jointly concave, and so is $g_u^s(\cdot)$ since it is a weighted aggregation of jointly concave functions.

For threshold-based loss functions where $t_u^s(\phi_u^s) = \alpha_u^s$ we have that

$$\begin{aligned} g_u^s(\phi_u^s, \gamma_u^s) &= \mathrm{E}[f_u^s(\phi_u^s, \gamma_u^s D)] \\ &= \hat{r}_u^s \phi_u^{\pi,s} P(\gamma_u^s D \leq \phi_u^{\pi,s} \alpha_s^u) \\ &= \hat{r}_u^s \phi_u^{\pi,s} F_D(\frac{\phi_u^{\pi,s} \alpha_s^u}{\gamma_u^s}). \end{aligned}$$

Now using the same result on the perspective functions of variables the result follows. The truncated Pareto case can be easily verified by taking derivatives.

D. Proof of Theorem 3

Clearly $\mathcal{C}^{H,\delta} \subset \hat{\mathcal{C}}^{H,\delta}$. We will show that $\mathbf{c} \in \hat{\mathcal{C}}^{H,\delta}$ then their exists $\boldsymbol{\pi} = (\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}) \in \Pi^{H,\delta}$ such that $\mathbf{c} \leq \mathbf{c}^{\boldsymbol{\pi}}$ from which it follows that $\hat{\mathcal{C}}^{H,\delta} \subset \mathcal{C}^{H,\delta}$.

Suppose $\mathbf{c} \in \hat{\mathcal{C}}^{H,\delta}$, then it can be represented as a convex combination of policies $\Pi^{H,\delta}$, in each channel state. For example suppose for simplicity that for that in channel state $s \in \mathcal{S}$ we have that $\lambda \in [0,1]$ one time shares between two policies $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ to achieve throughputs for $u \in \mathcal{U}$ given by

$$r_u^s = \lambda r_u^{\boldsymbol{\pi}_1,s} + (1-\lambda) r_u^{\boldsymbol{\pi}_2,s}.$$

Consider u we have

$$\begin{aligned} r_u^s &= \lambda r_u^{\boldsymbol{\pi}_1,s} + (1-\lambda) r_u^{\boldsymbol{\pi}_2,s} \\ &= \lambda g_u^s(\phi_u^{\boldsymbol{\pi}_1,s}, \gamma_u^{\boldsymbol{\pi}_1,s}) + (1-\lambda) g_u^s(\phi_u^{\boldsymbol{\pi}_2,s}, \gamma_u^{\boldsymbol{\pi}_2,s}) \\ &\leq g_u^s(\lambda \phi_u^{\boldsymbol{\pi}_1,s} + (1-\lambda)\phi_u^{\boldsymbol{\pi}_2,s},\ \lambda \gamma_u^{\boldsymbol{\pi}_1,s} + (1-\lambda)\phi_u^{\gamma_2,s}) \\ &= g_u^s(\phi_u^{\boldsymbol{\pi},s}, \gamma_u^{\boldsymbol{\pi},s}), \end{aligned}$$

where $\phi_u^{\boldsymbol{\pi},s} = \lambda \phi_u^{\boldsymbol{\pi}_1,s} + (1-\lambda)\phi_u^{\boldsymbol{\pi}_2,s}$ and $\gamma_u^{\boldsymbol{\pi},s} = \lambda \gamma_u^{\boldsymbol{\pi}_1,s} + (1-\lambda)\gamma_u^{\boldsymbol{\pi}_2,s}$. Clearly $\boldsymbol{\phi}^{\boldsymbol{\pi}}, \boldsymbol{\gamma}^{\boldsymbol{\pi}}$ as given above correspond to a policy $\boldsymbol{\pi}$ such that $\boldsymbol{\pi} \in \Pi^{H,\delta}$ since the set is convex. It also follows that $r_u^s \leq r_u^{\boldsymbol{\pi},s}$, so $c_u^s \leq c_u^{\boldsymbol{\pi},s}$ and so $\mathbf{c} \leq \mathbf{c}^{\boldsymbol{\pi}}$.

E. Proof of Theorem 4

The proof requires intermediate lemmas, detailed below. For the ease of exposition, let us define $U(\boldsymbol{r}) := \sum_{u\in\mathcal{U}} U_u(r_u)$ and $\nabla U(\boldsymbol{r}) := \left(\frac{\partial U_1(x)}{\partial x}\Big|_{x_1=r_1}, \frac{\partial U_2(x)}{\partial x}\Big|_{x_2=r_2}, \ldots, \frac{\partial U_1(x)}{\partial x}\Big|_{x_{|\mathcal{U}|}=r_{|\mathcal{U}|}}\right)^T$. First we have the following important lemma regarding the stochastic approximation algorithm.

Lemma 2. *$\boldsymbol{R}(t) = \left(R_1(t), R_2(t), \ldots, R_{|\mathcal{U}|}\right)^T$ is an unbiased estimator of argmax: $\nabla U\left(\overline{\boldsymbol{R}}(\boldsymbol{t})\right)^T \boldsymbol{c}$, i.e.,*

$$\mathbb{E}\left[\boldsymbol{R}(t)\right] = \underset{\boldsymbol{c}\in\mathcal{C}^{H,\delta}}{\textit{argmax:}}\ \nabla U\left(\overline{\boldsymbol{R}}(\boldsymbol{t})\right)^T \boldsymbol{c}. \tag{29}$$

Proof. Based on the definition of $\mathcal{C}^{H,\delta}$ we can re-write $\underset{\boldsymbol{c}\in\mathcal{C}^{H,\delta}}{\max:}\ \nabla U\left(\overline{\boldsymbol{R}}(\boldsymbol{t})\right)^T \boldsymbol{c}$ as follows:

$$\max_{\boldsymbol{\phi},\boldsymbol{\gamma}} \quad \sum_{u\in\mathcal{U}} U_u^{'}\left(\overline{R}_u(t)\right)\left(\sum_{s\in\mathcal{S}} p_{\mathcal{S}}(s) g_u^s\left(\phi_u^s, \gamma_u^s\right)\right), \tag{30}$$

$$\text{s.t.} \quad \boldsymbol{\phi} \geq (1-\delta)\,\boldsymbol{\gamma}, \tag{31}$$

$$\boldsymbol{\phi},\ \boldsymbol{\gamma} \in \Pi^{H,\delta}. \tag{32}$$

Observe that the above optimization problem can be solved separately for each network state $s \in \mathcal{S}$. The de-coupled problem for any state s is same as the optimization problem (9) in our online algorithm. With a slight abuse of notation, let $\left(\tilde{\boldsymbol{\phi}}(s), \tilde{\boldsymbol{\gamma}}(s)\right)$ be the optimal solution to the online problem when $S(t) = s$. Conditioned on $S(t) = s$, we have that:

$$\begin{aligned}\mathbb{E}\left[R_u(t) \mid S(t) = s\right] &= \mathbb{E}\left[f_u^s\left(\tilde{\phi}_u^s, \tilde{\gamma}_u^s D\right) \mid S(t) = s\right] \\ &= g_u^s\left(\tilde{\phi}_u^s, \tilde{\gamma}_u^s\right) \quad \forall u \in \mathcal{U}. \end{aligned} \tag{33}$$

Computing $\mathbb{E}\left[\mathbb{E}\left[R_u(t) \mid S(t)\right]\right]$ gives the desired result (29). □

The main intuition behind the proof of optimality is that for large t, the trajectories of $\overline{\boldsymbol{R}}(t)$ can be approximated by the solution to the following differential equation in $\boldsymbol{x}(t)$ with continuous time t:

$$\frac{d\boldsymbol{x}(t)}{dt} = \underset{\boldsymbol{c}\in\mathcal{C}^{\mathcal{H},\delta}}{\text{argmax:}}\ \nabla U\left(\boldsymbol{x}(t)\right)^T \boldsymbol{c} - \boldsymbol{x}(t). \tag{34}$$

Let us define $q(\boldsymbol{x}) := \underset{\boldsymbol{c}\in\mathcal{C}^{\mathcal{H},\delta}}{\text{argmax:}}\ \nabla U\left(\boldsymbol{x}\right)^T \boldsymbol{c}$. To show the optimality of our online algorithm, we shall also require the following result on the above differential equation.

Lemma 3. *The differential equation* (34) *is globally asymptotically stable. Furthermore, for any initial condition $\boldsymbol{x}(0) \in C^{\mathcal{H},\delta}$, we have that $\lim_{t\to\infty} \boldsymbol{x}(t) = \boldsymbol{r}^*$.*

Proof. To prove this lemma it is enough to show that there exists a Lyapunov function $L(\boldsymbol{x}(t))$ such that it has a negative drift when $x(t) \neq \boldsymbol{r}^*$ and has zero drift when $x(t) = \boldsymbol{r}^*$. Define $L(\boldsymbol{x}) = U(\boldsymbol{r}^*) - U(\boldsymbol{x})$. Observe that under our assumption of strictly concave $U_u(\cdot)$, the offline optimization problem is guaranteed to have an unique optimal solution, which is $\boldsymbol{r}^*$. Therefore, $\forall \boldsymbol{x} \in C^{\mathcal{H},\delta}$ and $\boldsymbol{x} \neq \boldsymbol{r}^*$ $L(\boldsymbol{x}) > 0$. Next we will compute the drift of $L(\boldsymbol{x}(t))$ with respect to time.

$$\frac{dL(\boldsymbol{x}(t))}{dt} = -\nabla U\left(\boldsymbol{x}(t)\right)^T \frac{d\boldsymbol{x}(t)}{dt}, \tag{35}$$

$$= -q\left(\boldsymbol{x}(t)\right) + \nabla U\left(\boldsymbol{x}(t)\right)^T \boldsymbol{x}(t), \tag{36}$$

$$< 0 \qquad \forall \boldsymbol{x}(t) \neq \boldsymbol{r}^*. \tag{37}$$

To get inequality (37), first observe that from the definition of $q(\boldsymbol{x}(t))$ and (36), we get that $\frac{dL(\boldsymbol{x}(t))}{dt} \leq 0$. However, we have to show that this inequality is strict for $\boldsymbol{x}(t) \neq \boldsymbol{r}^*$. Observe that $q(\boldsymbol{x}) = \boldsymbol{x}$ is a necessary and sufficient condition for optimality of the offline optimization problem, see [21] for more details. From strict concavity of the utility functions, we have an unique optimal point $\boldsymbol{r}^*$. Therefore, $\frac{dL(\boldsymbol{x}(t))}{dt} < 0$ for $\boldsymbol{x}(t) \neq \boldsymbol{r}^*$ and $\frac{dL(\boldsymbol{x}(t))}{dt} = 0$ at $\boldsymbol{x}(t) = \boldsymbol{r}^*$. □

To conclude the proof, Lemmas 2 and 3 along with the condition 3 satisfy all the conditions necessary to apply Theorem 2.1 in Chapter 5, [22] which states that $\overline{\boldsymbol{R}}(t)$ converges to $\boldsymbol{r}^*$ almost surely.

F. Proof of Theorem 5

The proof has the following two steps.

1) We shall first consider a hypothetical *non-casual* scenario and show that there exists an optimal joint scheduling policy with minislot-homogeneous URLLC placement policy which in general is a function of the aggregate URLLC load in an eMBB slot. We then upper bound the optimal value of $\mathcal{OP}_1$ by the solution to a hypothetical non-causal scenario described in the sequel.
2) Secondly, under Assumption 4 on the loss functions, we show that there exists an URLLC placement policy policy which is minislot-homogeneous but independent of the aggregate URLLC load for the hypothetical non-causal scenario. We then conclude that there exists an optimal minislot-homogeneous joint sceduling policy for $\mathcal{OP}_1$ as an upper bound for its value is attained by a minislot-homogeneous joint scheduling policy.

The two steps are elaborated next.

1) Hypothetical non-causal scenario: First let us describe the *non-causal* scenario. At the beginning of each eMBB slot, first the scheduler chooses $\phi^{\pi,s}$. Next the total URLLC demand in each minislot is revealed, i.e., the realizations of $D(1), D(2), \ldots, D(|\mathcal{M}|)$ are revealed. Therefore, this setting is not causal as it assumes knowledge about future URLLC demand realizations. In general the URLLC placement under the non-causal setting is dependent on the minislot index m and $\boldsymbol{D}^{(1:|\mathcal{M}|)}$. With slight abuse of notation, we shall denote it by $\gamma_{u,m}^s\left(\boldsymbol{D}^{(1:|\mathcal{M}|)}\right)$. The joint scheduling policy has to satisfy the constraints (15), (16), and (17). We have the following lemma on the *non-causal* setting.

Lemma 4. *There exists an optimal minislot-homogeneous policy for the non-casual setting such that the URLLC placement*

depends only on the total URLLC demand in an eMBB slot, i.e., $\sum_{m=1}^{|\mathcal{M}|} D_m$.

Proof. Let $\left(\tilde{\phi}^{\pi}, \tilde{\gamma}^{\pi,s}(\cdot)\right)$ be the decision variables under an optimal joint scheduling policy π in the non-causal setting. Let $d(1)$, $d(2), \ldots, d(|\mathcal{M}|)$ be realizations of $D(1), D(2), \ldots, D(|\mathcal{M}|)$ such that $\sum_{m=1}^{|\mathcal{M}|} d(m) = d$. Define the following:

$$\nu_u^s := \frac{\sum_{m=1}^{|\mathcal{M}|} \tilde{\gamma}_{u,m}^{\pi,s}\left(\boldsymbol{d}^{(1:|\mathcal{M}|)}\right) d(m)}{d}. \tag{38}$$

Note that with the definition of ν_u^s, the total puncturing experienced by an eMBB user u in an eMBB slot is $\nu_u^s d$. From this one can construct an equivalent minislot-homogeneous URLLC placement policy. For all minislots, use ν^s as the URLLC placement factor. This satisfies the constraints (15), (16), and (17). In general ν^s could depend on $d(1)$, $d(2)$, $\ldots$, $d(|\mathcal{M}|)$. However, we will show that the optimal solution depends only on the sum $\sum_{m=1}^{|\mathcal{M}|} d_m$.

Let $d'(1)$, $d'(2), \ldots,$ $d'(|\mathcal{M}|)$ be such that $\sum_{m=1}^{|\mathcal{M}|} d'_m = d$ and there exists an m such that $d'(m) \neq d(m)$. Define the following:

$$\nu_u'^s := \frac{\sum_{m=1}^{|\mathcal{M}|} \tilde{\gamma}_{u,m}^{\pi,s}\left(\boldsymbol{d}'^{(1:|\mathcal{M}|)}\right) d'_m}{d}. \tag{39}$$

Therefore, the total puncturing observed by $\nu_u'^s d$. Observe that ν'^s is also a feasible URLLC policy for the case when the URLLC demand realizations are $d(1)$, $d(2)$, $\ldots$, $d(|\mathcal{M}|)$. Similarly ν^s is also a feasible URLLC placement policy for the case with $d'(1)$, $d'(2)$, $\ldots$, $d'(|\mathcal{M}|)$. Therefore, the optimal solution has to be independent of the realizations of $D(1)$, $D(2), \ldots, D(|\mathcal{M}|)$ and depends only on the sum $\sum_{m=1}^{|\mathcal{M}|} D_m$. □

Therefore, we shall restrict ourselves to minislot-homogeneous policies in the non-causal setting with the URLLC placement as a function of the total URLLC demand for that eMBB slot. With slight abuse of notation we shall denote a URLLC placement policy in this setting by $\gamma_u^s(\cdot)$ with the only argument as the total URLLC demand in that eMBB slot. This procedure is formally described next.

1) At the beginning of an eMBB slot, the joint scheduler chooses $\phi_u^{\pi,s}, u \in \mathcal{U}$ such that

$$\sum_{u\in\mathcal{U}} \phi_u^{\pi,s} = 1 \text{ and } \phi_u^{\pi,s} \in [0,1] \quad \forall u. \tag{40}$$

2) The total URLLC demand $D = \sum_{m=1}^{|\mathcal{M}|} D(m)$ in that eMBB slot is revealed.
3) For an URLLC demand of D, $\gamma_u^{\pi,s}(D)$ is chosen such that

$$\sum_{u\in\mathcal{U}} \gamma_u^{\pi,s}(D) = 1, \quad \text{and} \quad \gamma_u^{\pi,s}(D) \in [0,1]. \tag{41}$$

Let us denote the feasible policies for this hypothetical non-causal scenario by $\Pi^{\dagger}$. $(\boldsymbol{\phi^{\pi,s}}, \boldsymbol{\gamma^{\pi,s}})$ is chosen as the solution to the following optimization problem.

$$\mathcal{OP}_2: \quad \max_{\pi\in\Pi^{\dagger}} : \sum_{u\in\mathcal{U}} w_u g_u^{\pi,s}\left(\phi_u^{\pi,s}, \gamma_u^{\pi,s}(\cdot)\right), \tag{42}$$

where $g_u^{\pi,s}\left(\phi_u^{\pi,s}, \gamma_u^{\pi,s}(\cdot)\right) = r_u^s \phi_u^{\pi,s} \mathbb{E}\left[1 - h_u^s \left(\frac{\gamma_u^{\pi,s}(D) D}{\phi_u^{\pi,s}}\right)\right]$. We have the following important lemma which states that the optimal value under the non-causal scenario is an upper bound to the optimal value under the causal and minislot-dependent policy.

Lemma 5.

$$\max_{\pi\in\Pi^{\dagger}} : \sum_{u\in\mathcal{U}} w_u g_u^{\pi,s}\left(\phi_u^{\pi,s}, \gamma_u^{\pi,s}(\cdot)\right) \geq \max_{\pi\in\tilde{\Pi}} : \sum_{u\in\mathcal{U}} w_u g_u^{\pi,s}\left(\phi_u^{\pi,s}, \boldsymbol{\gamma}_u^{\pi,s}\right). \tag{43}$$

Proof. This directly follows from the proof of Lemma 4 where we have shown that any URLLC placement factor $\boldsymbol{\gamma}_u^{\pi,s}$ can be transformed into a minislot-homogeneous policy which depend only on the total URLLC demand in an eMBB slot, and hence, any feasible solution for $\mathcal{OP}_1$ is a feasible solution for $\mathcal{OP}_2$. □

2) Existence of an optimal solution independent of the value of D: In general the optimal URLLC placement policy under $\mathcal{OP}_2$ may depend on the total URLLC demand in an eMBB slot. However, under the Assumption 4 it is independent of the total URLLC demand. This is stated formally in the following lemma.

Lemma 6. *Under Assumption 4, there exists an optimal solution* $(\boldsymbol{\phi}^{*,s}, \boldsymbol{\gamma}^{*,s}(\cdot))$ *for* $\mathcal{OP}_2$ *with URLLC placement policy* $(\boldsymbol{\gamma}^{*,s}(\cdot))$ *independent of* D.

Proof. If $(\boldsymbol{\phi}^{*,s}, \boldsymbol{\gamma}^{*,s}(\cdot))$ is an optimal solution to $\mathcal{OP}_2$, then $\boldsymbol{\gamma}^{*,s}(\cdot)$ must also be an optimal solution to the following optimization problem in $\boldsymbol{\gamma}^s := \left(\gamma_1^s(\cdot), \gamma_2^s(\cdot), \ldots, \gamma_{|\mathcal{U}|}^s(\cdot)\right)$.

$$\max_{\boldsymbol{\gamma}^s} \quad \sum_{u\in\mathcal{U}} w_u g_u^s(\phi_u^{*,s}, \gamma_u^s(\cdot)), \tag{44}$$

$$\text{s.t.} \quad \phi_u^{*,s} \geq (1-\delta)\gamma_u^s(d) \quad \forall u,\, d, \tag{45}$$

$$\sum_{u\in\mathcal{U}} \gamma_u^s(d) = 1 \text{ and } \gamma_u^s(d) \in [0,1] \quad \forall u,\, d. \tag{46}$$

$$\tag{47}$$

For any d and u, from the K.K.T. conditions for the above optimization problem, we have that

$$-w_u r_u^s d^p h_u^{s'} \left(\frac{\gamma_u^{*,s}(d)}{\phi_u^{*,s}}\right) + \beta(d) + \eta_u(d) - \nu_u(d) - \lambda_u(d) = 0. \tag{48}$$

where $h_u^{s'}(x) = \frac{dh_u^s(y)}{dy}\Big|_{y=x}$,$\beta(d)$ is an arbitrary constant (function of d) and $\eta_u(d)$, $\nu_u(d)$ and $\lambda_u(d)$ are constants such that

$$\lambda_u(d)\left(\phi_u^{*,s}(d) - \gamma_u^{*,s}(1-\delta)\right) = 0 \quad \text{and} \quad \lambda_u(d) \geq 0 \quad \forall u, \tag{49}$$

$$\eta_u(d)\gamma_u^{*,s}(d) = 0 \quad \text{and} \quad \eta_u(d) \geq 0 \quad \forall u, \tag{50}$$

$$\nu_u(d)\left(1 - \gamma_u^{*,s}(d)\right) = 0 \quad \text{and} \quad \nu_u(d) \geq 0 \quad \forall u. \tag{51}$$

Note that we have used the fact that for a homogeneous loss functions $h_u^{s'}(dx) = d^p h_u^{s'}(x)$. For any $\tilde{d} \neq d$, if we choose $\beta(\tilde{d}) = \beta(d)\frac{\tilde{d}^p}{d^p}$, $\eta_u(\tilde{d}) = \eta_u(d)\frac{\tilde{d}^p}{d^p}$, $\nu_u(\tilde{d}) = \nu_u(d)\frac{\tilde{d}^p}{d^p}$, and $\lambda_u(\tilde{d}) = \lambda_u(d)\frac{\tilde{d}^p}{d^p}$, then from (48) $\gamma_u^{*,s}(d)$ and $\phi_u^{*,s}$ satisfy the K.K.T. condition for $\tilde{d}$

$$-w_u r_u^s \tilde{d}^p h^{s'}\left(\frac{\gamma_u^{*,s}(d)}{\phi_u^{*,s}}\right) + \beta(\tilde{d}) + \eta_u(\tilde{d}) - \nu_u(\tilde{d}) - \lambda_u(\tilde{d}) = 0. \tag{52}$$

Hence, $\gamma_u^{*,s}(d)$ and ϕ_u^s are optimal for $\tilde{d}$ too. Hence, we have a constructed an optimal solution with URLLC placement policy independent of D. □

We have shown in Lemma 6 that there exists an optimal policy $(\boldsymbol{\phi}^{*,s}, \boldsymbol{\gamma}^{*,s})$ which is a minislot-homogeneous policy and independent of the realization of D. In Lemma 5, we have also shown that the optimal value of $\mathcal{OP}_2$ is an upper bound for $\mathcal{OP}_1$. Hence, there exists a minislot-homogeneous policy which achieves an upper bound for $\mathcal{OP}_1$. Therefore, there exists a minislot-homogeneous policy which is optimal for $\mathcal{OP}_1$.

G. Proof of Theorem 6

Let $\mathcal{S}_k$ be the set of all subsets with k elements chosen from the set $\{1, 2, \ldots, m_1 + m_2\}$. For example, if $m_1 + m_2 = 3$ and $k = 2$, then $\mathcal{S}_k = \{\{1,2\}, \{2,3\}, \{1,3\}\}$. Note that $|\mathcal{S}_k| = \binom{|\mathcal{M}|}{k}$. Using the above definitions, we can re-write the R.H.S. of (22) as follows:

$$\mathbb{E}\left[h_1^s\left(\sum_{m=1}^{m_1+m_2}\phi_1 D(m)\right)\right] = \mathbb{E}\left[h_1^s\left(\frac{1}{\binom{m_1+m_2}{m_1}}\sum_{q\in\mathcal{S}_{m_1}}\left(\sum_{m\in q} D(m)\right)\right)\right]. \tag{53}$$

Using the above expression one can apply Jensen's inequality on the R.H.S. of (22), we have that

$$\mathbb{E}\left[h_1^s\left(\sum_{m=1}^{m_1+m_2}\phi_1 D(m)\right)\right] \leq \frac{1}{\binom{m_1+m_2}{m_1}}\sum_{q\in\mathcal{S}_{m_1}}\mathbb{E}\left[h_1^s\left(\sum_{m\in q} D(m)\right)\right]. \tag{54}$$

Since D_m's are i.i.d. the R.H.S. of the above expression is same as the L.H.S. of (22). Hence, proved.

H. Proof of Theorem 7

Clearly the probability of loss depends on the minislot demands and the users thresholds. If one relaxes the sequential constraint on URLLC allocations, one can consider aggregating the the minislot demands and pooling together the users superposition/puncturing thresholds. The probability of loss for this relaxed system is simply the probability the demand exceeds the size of the superposition/puncturing pool, i.e., The probability of loss under the pooled resources is given by

$$P(D \geq \sum_{u\in\mathcal{U}} \phi_u^s t_u^s(\phi_u^s)).$$

This is clearly a lower bound for any placement policy. Note however that the threshold proportional strategy meets this bound from Corollary 3 (see Equation (25)) so it indeed minimizes the probability of loss on a given eMBB slot.