A PROBLEM SETTING

A.1 TRAFFIC MODEL

The traffic model consider here is as follows: users appear and disappear continuously and each belongs to a specific class. The characteristics of those classes describe statistically the traffic. Each class has the following attributes:

- Data size D: the size of data a user of that class asks for.
- Maximum Latency L: the maximum number time slots within which the user needs to successfully receive the packet of size D, so as to be satisfied.
- Importance α: used to dictate the scheduler to prioritize some classes, for users with more privileged contracts (SLAs) demanding better service and higher reliability.
- Arrival probability p: the probability a new user belonging to this class arrives in the system.

We denote $\mathcal{C}$ the set of classes. Every user entering the system, belongs to a class $c \in \mathcal{C}$ with probability p_c and is characterized by the tuple (D_c, L_c, α_c). We assume that a maximum number K of users can coexist per time slot. We also assume that a new user appears whenever a previous one has reached the maximum time it can be present in the system. For example, if a user appears at time $t = 1$, belonging to a class $c \in \mathcal{C}$ with $L_c = 4$, then even if it gets immediately and immediately successfully its requested packet of size D_c at $t = 1$, it will still remain in the system until a new user appears at $t = 5$ belonging to a class $c' \in \mathcal{C}$ with probability $p_{c'}$. Therefore, at every time slot a set U_t (with constant cardinality $|U_t| = K$) of users is observed, with some of them belonging to set $U_t^{act} \subseteq U_t$, with some demanding resources from the base station as being unsatisfied and some being already satisfied. The assumption of always K users is alleviated by adding the null class c_0 with $D_{c_0} = 0$, $\alpha_{c_0} = 0$ and $L_{c_0} > 0$. A "user" of the null class is equivalent to no user has appeared and the fact that the number of user can fluctuate over time is incorporated.

The rationale behind this specific traffic model is as follows: (i) it is a traffic model under which users with different strict data and latency requirements come and go, and is both quite generic and tractable enough to permit to benchmark (this does not apply to DRL since it is model-free) (ii) the traffic remains uninfluenced by the scheduler decisions. On the contrary when the common assumption that whenever a user is satisfied a new one arrives with some probability per time slot is made, then the scheduler performance affects the statistics of the traffic because at a given time interval, a scheduler with abundant resources will see more users than one with poor resources, since more resources means satisfying users earlier leading with that model to (statistically) more users appearing.

A.2 GEOMETRY, CHANNEL AND RATE MODEL

A.2.1 GEOMETRY

The users are assumed to be uniformly distributed within a concentric ring. Therefore the distance of a user u from the base station is a random variable with a probability density function: $f_{\mathrm{d}}(\mathrm{d}_u) = \frac{2\mathrm{d}_u}{\mathrm{d}_{max}^2 - \mathrm{d}_{min}^2}, \mathrm{d}_u \in [\mathrm{d}_{min}, \mathrm{d}_{max}]$. Furthermore we assume that the mobility of the users is not too high to change significantly within their limited time interval they are active and experience a big change in the power of the received signal. Consequently, their distances from the base station are kept constant. In contrast, the modification of the channel due to small scale fading (i.e. small scale mobility) is taken into account and described below.

A.2.2 CHANNEL

Multiple users can be served simultaneously and in this work we assume that they are allocated on different orthogonal frequency bands (avoiding interference between them). We also assume that they experience flat block fading and therefore every user has a constant channel gain for a given time slot and throughout all the available frequency band from which is served. Let a user u that appeared at time t_0, with channel gain at time t is $g_{u,t} = \frac{C_{pl}|h_{u,t}|^2}{\sigma_N^2} \mathrm{d}_u^{-n_{pl}}$ with n_{pl} being the pathloss

exponent, C_{pl} a constant to account for the constant losses and σ_N^2 is the noise power spectrum density. The distance d_u remains constant throughout out the lifespan of user u but there is a small scale Rayleigh fading changing in every time slot according to the Markovian model:

$$h_{u,t_0} \sim \mathcal{CN}(0,1)$$
$$h_{u,t} = \rho h_{u,t-1} + Z, \quad \text{with} \quad Z \sim \mathcal{CN}(0, 1-\rho^2), t > t_0$$

where $\mathcal{CN}(0,v)$ represents a circular complex normal distribution with zero mean and variance v. The parameter $\rho = J_0(2\pi f_d T_{slot}) \in [0,1]$ (Tan & Beaulieu, 2000) determines the time correlation of the channel where $J_0(\cdot)$ being the zeroth-order Bessel function of the first kind, f_d the maximum Doppler frequency (determined by the mobility of the users) and T_{slot} the slot duration. If $\rho = 0$ (high mobility at the small scale level), in every time slot the user has an independent realization of the fading distribution. If $\rho = 1$ (absence of mobility), the fading is constant throughout the user's lifespan.

A.2.3 RATE

We assume the Shannon rate formula is valid and that the base station operates on capacity level providing to user u at time t data equal to $w_{u,t}\log_2(1+g_{u,t}\mathrm{P}_{u,t})$, where $\mathrm{P}_{u,t}$ is the transmitted energy per channel use/symbol and $w_{u,t}$ the assigned bandwidth. An outage happens when the user's data requirement is higher than what the channel can support. For instance if we consider transmission at given time t_u to user at distance d_u from the base station, belonging to class $c \in \mathcal{C}$ with resources $(w_{u,t}, \mathrm{P}_{u,t})$ then the probability of failing to correctly decode its packet equals to:

$$\begin{aligned} P_u^{fail}(w_{u,t}, \mathrm{P}_{u,t}; \mathrm{d}_u) &= \mathbb{P}(w_{u,t}\log_2(1+g_{u,t}\mathrm{P}_{u,t}) < D_u | \mathrm{d}_u) = \mathbb{P}(|h_{u,t}|^2 < \zeta_{u,t}\mathrm{d}_u^{n_{pl}}) \\ &= 1 - e^{-\zeta_{u,t}\mathrm{d}_u^{n_{pl}}} \end{aligned} \tag{2}$$

with $\zeta_{u,t}=\dfrac{\sigma_N^2(2^{D_u/w_{u,t}}-1)}{C_{pl}\mathrm{P}_{u,t}}$. Now if the location d_u of this user is unknown by the scheduler (corresponds who will appear in the future), the error probability becomes

$$\begin{aligned} P_u^{fail}(w_{u,t}, \mathrm{P}_{u,t}) &= \mathbb{P}(w_{u,t}\log_2(1+g_{u,t}\mathrm{P}_{u,t}) < D_u) = \int_{\mathrm{d}_{min}}^{\mathrm{d}_{max}} P_u^{fail}(w_{u,t}, \mathrm{P}_{u,t}; \mathrm{d}) f_{\mathrm{d}}(\mathrm{d}) d\mathrm{d} \\ &= 1 - \frac{\Gamma(\frac{2}{n_{pl}}, \zeta_{u,t}\mathrm{d}_{min}^{n_{pl}}) - \Gamma(\frac{2}{n_{pl}}, \zeta_{u,t}\mathrm{d}_{max}^{n_{pl}})}{n_{pl}\zeta_{u,t}^{2/n_{pl}}(\mathrm{d}_{max}^2 - \mathrm{d}_{min}^2)/2} \end{aligned} \tag{3}$$

where $\Gamma(s,x) = \int_x^\infty t^{s-1}\,\mathrm{e}^{-t}\,\mathrm{d}t$ is the upper incomplete gamma function. For the sake of simplicity, we overloaded notation by allowing x in D_x, α_x, L_x to either denote a class x or a user x belonging to a class with those characteristics.

A.3 SCHEDULING PROCEDURE

The base station is called to appropriately use in every time-slot its energy and bandwidth resources to satisfy its users. We concentrate only on the bandwidth distribution, assuming no power adaptation and simplifying the base station job that spends a fixed amount of energy per channel use, i.e., $\mathrm{P}_{u,t} = \mathrm{P}, \forall u,t$. If the available bandwidth at the base station's disposal is W then the scheduler aims to find the $(w_{u_1,t}, w_{u_2,t}, ...) \in \mathbb{R}_{\geq 0}^{|U_t^{act}|}$ with $u_1, u_2, ... \in U_t^{act}$ such that

$$\sum_{u \in U_t^{act}} w_{u,t} \leq W, \quad \forall t$$

and maximize over the time horizon the accumulated reward for every satisfied user which is described by the following objective "gain-function":

$$G = \sum_t \sum_{u \in U_t^{act}} \alpha_u \mathbf{1}\{w_{u,t}\log_2(1+g_{u,t}\mathrm{P}) > D_u\}. \tag{4}$$

We stress out that a user u remains on the set U_t^{act} for a time interval less or equal to its maximum acceptable latency L_u. If not satisfied within that interval then he does not contribute positively to the objective G.

As implied by (4), the retransmission protocol adopted is Hybrid Automatic Repeat reQuest (HARQ-type I) with rate adaptation. If a user fails to correctly decode the received packet then this packet is ignored (no buffering at the receiver side) and the user waits until the base station sends again the same packet to try to decode. Finally, as stated previously, we consider two different CSI cases. In the first scenario, *statistical CSI*, only statistical properties of channel and location are known at current time t_c and the future, while, in the second scenario, *full CSI*, the exact value of the channels $h_{u,t_c}, \forall u \in U_{t_c}^{act}$ and the location of the users (and so $\mathrm{d}_u \forall u \in U_{t_c}^{act}$) are known at the current time t_c.

B SCHEDULING METHODS

B.1 FRANK WOLFE

In this section we deal with the case where the scheduler knows all the statistical properties of the system, i.e. channel and traffic. We write the problem in an optimization form on which the Frank Wolfe method is applied to decide the resource allocation.

Let first concentrate on the case of a single user u_0 appearing at time t_0. The current time is $t_c \in [t_0, t_0 + L_{u_0} - 1]$. We denote by $\overrightarrow{w}_{u_0,t} = (w_{u_0,t_0}, w_{u_0,t_0+1}, ..., w_{u_0,t})$ the assigned bandwidth from time t_0 (beginning of transmission for user u_0). Additionally, let $A_{u_0,t}$ be a binary random variable which if $A_{u_0,t} = 1$ then u_0 is still unsatisfied at the end of time slot t (*after* receiving $\overrightarrow{w}_{u_0,t}$ resources) and $A_{u_0,t} = 0$ otherwise. Given that at the beginning of time t user u_0 is still unsatisfied and that we know the resource allocation $w_{u_0,t}$ is scheduled to be done at time t, we define $\Phi(\overrightarrow{w}_{u_0,t}; d_{u_0})$ to be the probability that $w_{u_0,t}$ is still not enough when the location d_{u_0} is known but the channel $h_{u_0,t}$ is unknown:

$$\Phi(\overrightarrow{w}_{u_0,t}; \mathrm{d}_{u_0}) = \begin{cases} \mathbb{P}(A_{u_0,t} = 1 | \overrightarrow{w}_{u_0,t-1}, \mathrm{d}_{u_0}, A_{u_0,t-1}{=}1), & t > t_0 \\ \mathbb{P}(A_{u_0,t} = 1 | \mathrm{d}_{u_0}), & t = t_c = t_0. \end{cases} \tag{5}$$

The average contribution of user u_0 to the gain function (4) on the time interval $[t_c, t]$ is given by the following equation, derived by applying the chain rule for conditional probability:

$$\mathfrak{g}_{u_0}^{[t_c,t]}{=}\mathfrak{g}(w_{u_0,t_c}, ..., w_{u_0,t}; \mathrm{d}_{u_0}) = \begin{cases} 0, & \text{if } t_c > t_0 \text{ and } A_{u_0,t_c-1}{=}0 \\ \alpha_{u_0}\big(1 - \prod_{j=t_c}^{t} \Phi(\overrightarrow{w}_{u_0,j}; \mathrm{d}_{u_0})\big), & \text{else.} \end{cases} \tag{6}$$

Now we consider that the average contribution on the gain function (4) for the the future users following the user u_0. The next user (if it exists) appears at time $t_1 = t_0 + L_{u_0}$, and so on. Therefore we consider the users noted as $u_1, u_2, \ldots$ that will appear at $t_1 = t_0 + L_{u_0}, t_2 = t_1 + L_{u_1}, \ldots$. We denote that with probabilities $p_{c_1}, p_{c_2}, \ldots$ they will belong to classes $c_1, c_2, \ldots$, respectively (and one of these classes may be the null class). These classes will determine the maximum latencies $L_{u_1}, L_{u_2}, \ldots$ and consequently the time arrivals $t_1, t_2, \ldots$ all being random variables. As we consider here future users, even their locations are unknown. Consequently we need to average over the locations the equations (5) and (6) to obtain their contribution on the gain function (4). So for $i \geq 1$ if $\overrightarrow{w}_{u_i,t} = (w_{u_i,t_i}, w_{u_i,t_i+1}, ..., w_{u_i,t})$, we have

$$\mathfrak{g}_{u_i}^{[t_i,t]}{=}\mathfrak{g}(w_{u_i,t_i}, ..., w_{u_i,t}) = \alpha_{u_0}\big(1 - \prod_{i=t_c}^{t} \Phi(\overrightarrow{w}_{u_i,i})\big) \tag{7}$$

where the contribution looking at time t with $t < t_i + L_{u_i}$ starts at time t_i for user u_i and where

$$\Phi(\overrightarrow{w}_{u_i,t}) = \begin{cases} \mathbb{P}(A_{u_i,t} = 1 | \overrightarrow{w}_{u_i,t-1}, A_{u_i,t-1}{=}1), & t > t_i \\ \mathbb{P}(A_{u_i,t} = 1), & t = t_i. \end{cases} \tag{8}$$

Hence, the averaged value of gain function for the sequence of users $u_0, u_1, ...$ (so when one user at most is active per time slot, i.e. $K = 1$) starting at the current time t_c is:

$$\mathcal{G}(w_{u_0,t_c}, ..., w_{u_0,t_1-1}, w_{u_1,t_1}, ...) =$$
$$\mathfrak{g}_{u_0}^{[t_c,t_1-1]}(.; d_{u_0}) + \sum_{c_1 \in \mathcal{C}} \left(p_{c_1} \cdot \mathfrak{g}_{u_1}^{[t_1,t_2-1]}(.) + \sum_{c_2 \in \mathcal{C}} \left(p_{c_2} \cdot \mathfrak{g}_{u_2}^{[t_2,t_3-1]}5.) + \sum_{c_3 \in \mathcal{C}} (...) \right) \right). \tag{9}$$

From (9), we observe a tree structure[3] that when a user vanishes there is a summation over all the possibilities of the classes that the new user can belong to. Therefore a number of branches is equal to the number of possible classes ($|\mathcal{C}|$). To manage the scalability issue, we cut the tree by considering only T future time slots and work with the finite horizon $[t_c, t_c + T - 1]$.

Finally, the general case with multiple users served simultaneously ($K > 1$) is easy to be considered by just computing K "parallel trees". With a slight abuse of notation, we consider that the first subscript of the variables w now refers to the index of the tree (and implicitly to a specific user). As a consequence, the variables for the scheduled bandwidth resources over an horizon of length T can be put into the following matrix:

$$\mathbf{W}_{t_c} = \begin{bmatrix} w_{1,t_c} & w_{1,t_c+1} & \cdots & w_{1,t_c+T-1} \\ w_{2,t_c} & w_{2,t_c+1} & \cdots & w_{2,t_c+T-1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K,t_c} & w_{K,t_c+1} & \cdots & w_{K,t_c+T-1} \end{bmatrix}$$

and the average gain for these resources takes the following form:

$$G(\mathbf{W}_{t_c}) = \sum_{k=1}^{K} \mathcal{G}(w_{k,t_c}, w_{k,t_c+1}, \cdots, w_{k,t_c+T-1}). \tag{10}$$

Finally we arrive at our optimization problem whose solution constitutes the benchmark procedure for the statistical CSI case is the following one at current time t_c:

$$\max_{\mathbf{W}_{t_c} \in \mathbb{R}_{\geq 0}^{K \times T}} \quad G(\mathbf{W}_{t_c}) \tag{11}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} w_{k,t} \leq W, \quad \forall t \in \{t_c, \ldots, t_c+T-1\}. \tag{12}$$

It can be easily shown that the objective function $G(\cdot)$ is non-concave with multiple local optimums. In contrast, the constraints given by equation (12) describe a compact and convex domain set which allows the application of so-called Frank-Wolfe algorithm (Frank & Wolfe, 1956). The idea behind this algorithm is as follows: at each iteration, the algorithm starts from a point and approximates the objective function around it with a linear (first-order) approximation. Then it solves the corresponding Linear Programming problem (LP) to find the best solution which will be the starting point of the next iteration. The procedure terminates when the algorithm converges to a local optimum, i.e., when the objective function does not increase anymore significantly. In order to exhibit a solution close to the global optimum, the algorithm is repeated N_{init} times with different randomly chosen initial points. At the end, we peak the best local optimum.

We provide some general remarks:

- The above benchmark procedure takes into account the past through (5) since all the previously allocated resources are involved.
- The procedure at current time t_c proposes a solution for the scheduler for both the current time t_c and for the future $[t_c+1, t_c+T-1]$. Nevertheless, as this procedure will be recomputed at time $t_c + 1$ (once the actions proposed for time t_c is applied and new information about the transmission's success or failure are available), the actions proposed at time t_c for time $t_c + 1$ are generally not applied. Obviously we will apply at time $t_c + 1$ the solution advocated by the procedure computed at time $t_c + 1$.
- The Frank-Wolfe method is `sublinear but` the computation of the objective function (10) and its partial derivatives grow `exponentially` with T which leads in practice to a slow and cumbersome method if one wants to account for the impact of distant future (not to mention that to be sure to retrieve a good local optimum we have to repeat the process N_{init} times).
- Lastly, the algorithm treats the *"mean"* case. It does not specify what really happens in the future since it only evaluate what happens in the future on average (for example over

[3] A simple way to be computed is recursively

every possible class of future user). It would be possible to address for every future scenario differently but by skyrocketing the number of variables and constraints, making the already-slow benchmark procedure slower.

B.1.1 CALCULATING THE EXPRESSIONS

Hereafter, we concentrate on calculating (5) and (8) for different channel model subcases. The rest of the benchmark procedure is straightforward[4].

• *The i.i.d. fading channel case ($\rho = 0$)*: It is the simplest subcase since there are no time dependencies on the fading, so equations (5) and (8) become

$$\Phi(\overrightarrow{w}_{u_0,t}; \mathrm{d}_{u_0}) = P_{u_0}^{fail}(w_{u_0,t}, P; \mathrm{d}_{u_0}), and \tag{13}$$

$$\Phi(\overrightarrow{w}_{u_i,t}) = P_{u_i}^{fail}(w_{u_i,t}, P), \quad i \geq 1. \tag{14}$$

We remind the users u_i for $i \geq 1$ follow the user u_0, and therefore we average over their unknown locations.

• *The constant fading channel case ($\rho = 1$)*: Now the channel is the same for each retransmission on the user. For user u_0, the channel is invariant but unknown. Only its location is known. At time $t > t_0$, we have

$$\begin{aligned}\Phi(\overrightarrow{w}_{u_0,t}; \mathrm{d}_{u_0}) &= \mathbb{P}(w_{u_0,t}\log(1+g_{u_0}\mathrm{P}) < D_{u_0} | w_{u_0,t'}\log(1+g_{u_0}\mathrm{P}) < D_{u_0}, \forall t' \in [t_0, t-1], d_{u_0}) \\ &= \frac{\mathbb{P}(w_{u_0,t''}\log(1+g_{u_0}\mathrm{P}) < D_{u_0}, \forall t'' \in [t_0, t] | d_{u_0})}{\mathbb{P}(w_{u_0,t'}\log(1+g_{u_0}\mathrm{P}) < D_{u_0}, \forall t' \in [t_0, t-1] | d_{u_0})}.\end{aligned}$$

Therefore we obtain

$$\Phi(\overrightarrow{w}_{u_0,t}; \mathrm{d}_{u_0}) = \begin{cases} \dfrac{P_{u_0}^{fail}(\max\{\overrightarrow{w}_{u_0,t}\}, P; \mathrm{d}_{u_0})}{P_{u_0}^{fail}(\max\{\overrightarrow{w}_{u_0,t-1}\}, P; \mathrm{d}_{u_0})}, & \text{if } t > t_0 \\ P_{u_0}^{fail}(w_{u_0,t}, P; \mathrm{d}_{u_0}), & \text{if } \quad t = t_0. \end{cases} \tag{15}$$

For the case of the future users (u_i with $i \geq 1$), the equations remain the same with the only change that the location of the users is unknown as well. So in equation (15), we just need to omit the d_u similarly to the i.i.d. case.

• *The general Markovian case ($\rho \in (0, 1)$)*: Let us focus on the user u_0 and we are looking at the time $t = t_0 + 1$. According to (Nuttall, 1975, eq: 37), we have:

$$\begin{aligned}\Phi(\overrightarrow{w}_{u_0,t_0+1}; \mathrm{d}_{u_0}) &= \int_0^{x_{u_0,0}} \int_0^{x_{u_0,1}} \mathbb{P}(|h_{u_0,t_0+1}| = x\,|y)\mathbb{P}(|h_{u_0,t_0}| = y) dx dy \\ &= 1 - \frac{e^{-x_1^2} Q_1(\frac{x_{u_0,0}}{\sigma_R}, \frac{\rho x_{u_0,1}}{\sigma_R}) - e^{-x_{u_0,0}^2} Q_1(\frac{\rho x_{u_0,0}}{\sigma_R}, \frac{x_{u_0,1}}{\sigma_R})}{2(1 - e^{-x_{u_0,0}^2})}\end{aligned} \tag{16}$$

with $x_{u_i,j} = \sqrt{\zeta_{u_i,t_i+j}} d^{-\frac{n_{pl}}{2}}, i \in \{0, 1\}$ and Q_M be the Marcum Q-function.
For the future users ($u_i, i \geq 1$), we have at time $t = t_i + 1$ (we remind that user u_i starts its transmission at time t_i):

$$\Phi(\overrightarrow{w}_{u,t_i+1}) = \int_{\mathrm{d}_{min}}^{\mathrm{d}_{max}} \Phi(\overrightarrow{w}_{u_i,t_i+1}; \mathrm{d}_{u_i}) f_{\mathrm{d}}(\mathrm{d}) d\mathrm{d}. \tag{17}$$

where $\Phi(\overrightarrow{w}_{u_i,t_i+1}; \mathrm{d}_{u_i})$ is given by (16) by replacing u_0 with u_i. This equation (17) is already *intractable* whereas we are just focusing on the two first adjacent retransmissions. Obviously, it is even worse if we consider more retransmissions. Therefore the benchmark procedure will be only designed for $\rho = 0$ or $\rho = 1$, even if tested in the general case $\rho \in (0, 1)$. More precisely, *for any ρ, we apply the benchmark procedure designed for either $\rho = 0$ or $\rho = 1$, and keep the best result.*

[4]Perhaps it is tricky to also find the derivative of (3) which is required for the first-order approximation in the Franck-Wolfe algorithm. So we get

$$\frac{dP_u^{fail}}{dw} = \int_{\mathrm{d}_{min}}^{\mathrm{d}_{max}} \frac{d\mathbb{P}(|h|^2 < \zeta_{u,t}\mathrm{d}^{n_{pl}})}{d\zeta_{u,t}} f_{\mathrm{d}}(\mathrm{d}) d\mathrm{d} \frac{d\zeta_{u,t}}{dw} = \frac{\Gamma(\frac{2+n_{pl}}{n_{pl}}, \zeta_{u,t}\mathrm{d}_{min}^{n_{pl}}) - \Gamma(\frac{2+n_{pl}}{n_{pl}}, \zeta_{u,t}\mathrm{d}_{max}^{n_{pl}})}{n_{pl}\zeta_{u,t}^{(2+n_{pl})/n_{pl}}(\mathrm{d}_{max}^2 - \mathrm{d}_{min}^2)/2} \frac{d\zeta_{u,t}}{dw}$$

This case is much more complicated due to the correlation between the channel realizations. Actually, at time t, the distribution of $h_{u,t}$ given the past (which is not known in practice) is Ricean distributed. More precisely, if the user u is active at $t-1$ and t, we have $\mathbb{P}(|h_{u,t}|{=}x \,\Big|\, |h_{u,t-1}|) = \mathrm{Rice}(x; v_R = \rho|h_{u_0,t-1}|, \sigma_R^2 = \frac{1-\rho^2}{2})$ where v_R and σ_R^2 are the so-called Ricean parameters.

B.2 Knapsack and Integer Linear Programming

In this section we deal with the case where the scheduler has full-CSI, i.e. knows the complete state of the system/environment. Let first work on the user u_0 at the current time $t_c (t_c \geq t_0)$ for which both the channel h_{u_0,t_c} and location d_{u_0} is known. But the future channels $h_{u_0,t}$ for $t > t_c$ are only statistically known. The user u_0 is unsatisfied at t iff the allocated bandwidth $w_{u_0,t}$ is smaller than the following threshold

$$w_{u_0,t}^{th} = \frac{D_{u_0}}{\log_2(1 + g_{u_0,t}\mathrm{P})}.$$

Consequently, the error probability of user u_0 defined by (5) can be expressed:

$$\Phi(\overrightarrow{w}_{u_0,t}; \mathrm{d}_{u_0}) = \begin{cases} \mathbb{P}(w_{u_0,t} < w_{u_0,t}^{th} | A_{u_0,t-1} = 1, h_{u_0,t_c}, d_{u_0}), & \text{if } t > t_c \\ \mathbf{1}\{w_{u_0,t_c} < w_{u_0,t_c}^{th}\}, & \text{if } \quad t = t_c. \end{cases} \tag{18}$$

In this case, we remark that the probabilities are not necessary continuous due the indicator function in (18). Consequently, the gain function described in (10) is now non-continuous over the variables $w_{k,t_c} \forall k$ (because we know exactly the channel gains at t_c and indicator functions occur at this time), but continuous for $w_{k,t}, t > t_c$ corresponding to the future. To overcome this problem, we split the problem into two cases;

- Immediate horizon ($T = 1$): we focus only on the current time t_c and the effects on the future are omitted. We reach to a ***Knapsack*** which is myopically optimal.
- Finite horizon ($T > 1$): we take into account the future but we assume the channel realization and the location at time $t \in [t_c, t_c + T - 1]$ are known in advance, i.e., when the algorithm is run at time t_c. We write the problem as an ***integer linear programming (ILP)*** and the algorithm functions as an oracle providing an ***upper bound***.

B.2.1 Immediate horizon: $T = 1$

In this case, the optimization problem can be entirely restated. The variables to be optimized are x_{u,t_c} which is 1 if user u is active at time t_c or 0 otherwise. The cost in bandwidth is $w_{u,t_c}^{th} x_{u,t_c}$ because we assume that if an user is active, then the scheduler provides to it the minimum bandwidth it required to do a transmission without failure. Then the contribution in the gain function is $\alpha_u x_{u,t_c}$. Therefore the optimization problem can written as follows

$$\begin{aligned} \max_{x_{u,t_c}} \quad & \sum_{u \in U_{t_c}^{act}} \alpha_u x_{u,t_c} \\ \text{s.t.} \quad & \sum_{u \in U_{t_c}^{act}} w_{u,t_c}^{th} x_{u,t_c} \leq W \\ & x_{u,t_c} \in \{0,1\}, \quad \forall u \in U_{t_c}^{act}. \end{aligned}$$

This problem is a ***Knapsack*** problem, which corresponds to maximizing the total value by choosing from a set of objects a proper subset. Every object has its value but also a weight that prevents from picking all of them since the total weight of the chosen subset should not overreach the capacity level. It is a well known $\mathcal{NP}$-complete problem with various efficient algorithms for solving it and we used the library OR-TOOLS of Google.

B.2.2 Finite horizon: $T > 1$

As remarked previously, the original problem described by (18) is mixed, i.e. discrete over some variables and continuous over others. One idea is to approximate the indicator function with a

continuous function[5] in order to apply the Frank-Wolfe algorithm again as in the case of statistical CSI. We do not follow this way since the number of bad local optimums grow up and also it is very dependent on the choice of the approximating function. Hereafter, we assume that for the future $T-1$ time slots, the base station knows exactly how many and where users will appear, of which class and what will be their channels. The base station thus acts as an *oracle* capable to perfectly calibrate the scheduling to future fluctuations. We obtain therefore an upper bound of the performance of our policies.

Connecting this problem with a knapsack problem is not successful. Let us assume in the time interval $[t_c, t_c+T-1]$ the oracle knows a set of $U_{t_c}^T$ users/objects appear in total. We can think of having T different knapsacks (one for each $t \in [t_c, t_c+T-1]$ and all of capacity W), which we aim to fill with users/objects from the set $U_{t_c}^T$. The goal is to maximize the overall value of the chosen objects, i.e. satisfied users. This corresponds to a "multiple knapsack problem" but with a crucial difference. In contrast to "multiple knapsack problem", the weight of each object/user fluctuates over time as a consequence of the channel variability which changes the required resources/weight. That means that every object has a different weight depending on the knapsack it will be put it. Even considering $\rho = 1$, the constant channel does not help much since for some time slots in $[t_c, t_c+T-1]$ a user can happen to be either "unborn" or "dead". In those time slots we have to assume a different weight at those time slots that will be something greater than W so as to make it impossible to fit in the knapsacks corresponding to those time slots.

Finally we address our problem using a more generic (and slower) approach after formulating it as a ***Integer Linear Programming***. As mentioned, inside the lifespan $t \in I_{life} = [\max(t_c, t_u), \min(t_u+L_u-1, t_c+T-1)]$ of a user $u \in U_{t_c}^T$, $w_{u,t}^{th}$ are the (accurately predicted by the oracle) required bandwidth to satisfy u at time t given his channel gain $g_{u,t}$. Outside $t \in [t_c, t_c+T-1]/I_{life}$, $w_{u,t}^{th}$ is given a value greater than W so as to prevent any allocation. The formulation is

$$
\begin{aligned}
\max_{x_{u,t}} \quad & \sum_{u \in U_{t_c}^T} \alpha_u \sum_{t=t_c}^{t_c+T-1} x_{u,t} \\
\text{s.t.} \quad & \sum_{U_{t_c}^T} w_{u,t}^{th} x_{u,t} \leq W, \quad \forall t \in [t_c, t_c+T-1] \\
& \sum_{t=t_c}^{t_c+T-1} x_{u,t} \leq 1, \quad \forall u \in U_{t_c}^T \\
& x_{u,t} \in \{0,1\}, \quad \forall t \in [t_c, t_c+T-1] \text{ and } \forall u \in U_{t_c}^T.
\end{aligned}
$$

To solve this ILP optimization for every time step, we used the software CPLEX of IBM which relies on the Branch and Cut algorithm (Mitchell, 2002).

C PARAMETERS OF THE SETTING

The distance dependent path loss is set to be $120.9 + 37.6 \log_{10}(\mathrm{d})$ in dB which is compliant to LTE standard (LTE, 2018) and in our setting it translates to the constant loss component $C_{pl} = 10^{-12.09}$ and path loss exponent $n_{pl} = 3.76$. The AWGN power is $\sigma_N^2 = -149\text{dBm/Hz}$ (see appendix A.2).

For the DRL model we softly update the target policy and value network with momentum 0.005. We use replay buffer of capacity 5000 samples. The batch size is 64 and the learning rates equal to 0.001. The discount factor is set to $\gamma = 0.95$. With constant probability of 0.25 we explore according to what described in 3.2.1. We use $N_Q = 50$ of quantiles to describe the distribution. The ϕ_{user} is consisted of to fully connected layers with the hidden to be of dimension 10 and also its output. The input/output channels ratio of both f_{relu} and f_{linear} is 10/10. We remark that the number of parameters is relatively low (around 1000). We tried to increase but due to the high variance of the environment overfitting could not be avoided. Keeping also low the number of parameters makes it fast and cheap (both for energy and hardware) solution for practical use.

[5] So, the form $\mathbf{1}\{w > w_{u,t_c}^{th}\}$ needs to be changed into continuous function for which when $w < w_{u,t_c}^{th}$ it is equal to 0 in order to avoid giving less than w_{u,t_c}^{th} resource at user u and then it goes as fast as possible to 1.