Data Size-Aware Downlink Massive MIMO: A Session-Based Approach

Tung T. Vu, Quoc Ngo Hien, Minh N. Dao, Michail Matthaiou and Erik G. Larsson

The self-archived postprint version of this journal article is available at Linköping University Institutional Repository (DiVA):
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-187888

N.B.: When citing this work, cite the original publication.
https://doi.org/10.1109/LWC.2022.3174829

Original publication available at:
https://doi.org/10.1109/LWC.2022.3174829
Copyright: Institute of Electrical and Electronics Engineers
http://www.ieee.org/index.html
©2022 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Abstract—This letter considers the development of transmission strategies for the downlink of massive multiple-input multiple-output networks, with the objective of minimizing the completion time of the transmission. Specifically, we introduce a session-based scheme that splits time into sessions and allocates different rates in different sessions for the different users. In each session, one user is selected to complete its transmission and will not join subsequent sessions, which results in successively lower levels of interference when moving from one session to the next. An algorithm is developed to assign users and allocate transmit power that minimizes the completion time. Numerical results show that our proposed session-based scheme significantly outperforms conventional non-session-based schemes.

Index Terms—Massive MIMO, session-based, zero-forcing.

I. INTRODUCTION

We are witnessing an explosion of streaming and learning applications, such as video streaming, live conferencing, and federated learning [1]–[3]. Many of these applications require computations by mobile users (UEs) [4], and the UEs have fixed amounts of data to receive. To support these applications, it is critical to design transmission schemes that achieve low latency. It is of particular importance to design communication protocols that minimize the completion time, defined as the time it takes for a UE to receive all data destined for it.

To support the aforementioned applications, massive multiple-input multiple-output (MIMO) can be used due to its ability to offer high data rates to all UEs simultaneously [5]. To reduce the completion time, conventionally, the achievable rates of all the UEs are maximized via power allocation, and these rates are kept constant during the whole transmission. Another approach is to use different rates during the transmission period [6]. When some UEs have already completed their transmissions, other UEs will benefit by having less multi-user interference and higher rates, which results in shorter completion times. In this context, [6] studies the completion time for two-user systems from an information-theoretic perspective. However, general schemes for multi-user systems that use different rates within the transmission period have not been explored in the literature.

Contributions: Motivated by [6], we introduce a session-based scheme to reduce the completion times of UEs for the downlink of massive MIMO networks. In this scheme, UE data is transmitted with different rates in different sessions. Specifically, in each session, UEs are assigned so that one UE finishes its transmission and does not participate in subsequent sessions. UEs with uncompleted transmissions are allocated more power to obtain higher achievable rates, and complete their transmissions faster. Herein, zero-forcing (ZF) processing is used for data transmission. An algorithm is developed for assigning UEs and allocating transmit power, with the objective of minimizing the completion times of the UEs. Numerical results show that the proposed session-based scheme significantly reduces the completion times compared to conventional transmission that relies on power control only.

A specific version of this session-based scheme was also used in [7], although for a different objective and for a particular application. Herein, we substantially extend [7] to the general problem of minimizing downlink completion times. If the amounts of UE data in each session are given, then the optimization problem reduces to power and rate control for conventional transmission (with different rate constraints for different users). However, optimizing the amounts of data per session and the thresholds for the rate constraints over multiple sessions is a new and challenging problem.

II. SYSTEM MODEL

We consider the downlink transmission in a massive MIMO network, where an $M$-antenna base station (BS) serves $K \leq M$ single-antenna UEs simultaneously in the same frequency band. Let $S_k$ be the size of the data intended for UE $k$. We focus on applications where $S_k$ is fixed, such as mobile edge computing and federated learning to name but a few [3], [8]. We assume that the transmission time is within one large-scale coherence time, and the transmission is spans multiple small-scale coherence blocks. A small-scale coherence block is the time-frequency interval over which the channel is substantially static, and is divided into two phases: channel estimation and downlink payload data transmission.

Resource allocation, such as transmit power control, is typically performed to guarantee given quality-of-service targets. We consider two conventional schemes: (i) Conventional non-data size-aware scheme: This scheme allocates power such that all users achieve the same rate [5]; and (ii) Conventional data size-aware scheme: This scheme allocates power such that each UE receives a rate proportional to the size of its data. This is the traditional scheme used in wireless networks supporting mobile edge computing or federated learning (see, e.g., [3], [8] and references therein).

In both conventional schemes, the data rates are kept fixed for the whole transmission. However, since $K$ UEs have different required data sizes, some UEs could complete their transmissions before other UEs. Therefore, optimally, the data rates vary temporally depending on how many active UEs remain in the

$^1$The large-scale coherence time is the time during which the large-scale fading coefficients remain substantially constant [3].
Our proposed scheme uses different rates in different time periods for data transmissions. The transmission period during which the rates are kept fixed is called a “session”. More precisely, session \( i \) is defined as follows: (i) in session \( i \), there are \( K - i + 1 \) active UEs; and (ii) at the beginning of session \( i \), the BS updates the rates for these active UEs. These rates will be kept fixed until the end of this session, where one UE completes receiving data from the BS. Thus, the BS transmits data to all \( K \) UEs during \( K \) sessions. Fig. 1 illustrates the proposed session-based transmission as well as conventional transmission for a system with three UEs. Denote by \( a_i \) the indicator that is defined as
\[
a_{k,i} \triangleq \begin{cases} 
1, & \text{if UE } k \text{ is receiving data in session } i, \\
0, & \text{otherwise}. 
\end{cases} 
\]

Denote by \( K_i \triangleq \{ k | a_{k,i} = 1 \} \) the set of \( K_i = \sum_{k \in K} a_{k,i} \) UEs assigned in session \( i \in K \), where \( K = \{1, 2, \ldots, K\} \). Then, we have
\[
a_{k,1} = 1, \sum_{k \in K} a_{k,i} = K - i + 1, a_{k,i} \leq a_{k,i-1}, \forall i, \tag{2}
\]
to ensure that all the UEs will be served in session 1. In each subsequent session, one UE completes its transmission and will not join the next sessions. As such, more power is allocated to the UEs that have not yet completed their transmissions. Note that, the conventional (non-session-based) schemes are special cases of the proposed scheme when all UEs are served in session 1 with \( a_{k,1} = 1, \forall k \).

**Uplink channel estimation:** In each small-scale coherence block of length \( \tau_c \), each UE sends its pilot of length \( \tau_p \) to the BS. We assume that all pilots are mutually orthogonal, which requires \( \tau_p \geq K \).\(^2\) Denote by \( \mathbf{g}_k = (\beta_k)^{1/2} \mathbf{g}_0 \) the channel vector from UE \( k \) to the BS, where \( \beta_k \) and \( \mathbf{g}_0 \) are the large-scale fading coefficient and small-scale fading vector, respectively. With minimum mean-square error (MMSE) estimation, the channel estimate \( \hat{\mathbf{g}}_k \) of \( \mathbf{g}_k \) is distributed according to \( \mathcal{CN}(0, \sigma_k^2 \mathbf{I}_M) \), where \( \sigma_k^2 = \tau_p \rho_p \beta_k \), and \( \rho_p \) is the normalized transmit power of each pilot symbol \( [5, (3.8)] \). Let \( \mathbf{G}_i \triangleq [\mathbf{g}_1, \ldots, \mathbf{g}_K]_{i-1}, \forall k \in K_i \), be the matrix stacking the estimated channels of all participated UEs in session \( i \).

**Downlink payload data transmission:** In session \( i \), the BS uses ZF to transmit data to \( K - i + 1 \) UEs. With ZF, the signal transmitted by the BS is given by \( \mathbf{x}_i = \sqrt{\rho} \sum_{k \in K_i} \mathbf{u}_{k,i} s_{k,i} \), where \( \rho \) is the normalized transmit power at the BS; \( s_{k,i} \in \mathbb{C} \) and \( \mathbb{E}\{ |s_{k,i}|^2 \} = 1 \), is the symbol intended for UE \( k \); and
\[
\mathbf{u}_{k,i} = \sqrt{\eta_{k,i}} e_k \bar{G}_k (M - K_i) \bar{G}_k^H \bar{G}_i \bar{G}_i^H \mathbf{e}_k K_i \text{ is the ZF precoding vector. Here, } \mathbb{E}\{x\} \text{ denotes the expected value of a random variable } x, \text{ and } \mathbf{X}^H \text{ represents the conjugate transpose of a matrix } \mathbf{X}. \text{ In the precoding vector, } \eta_{k,i} \text{ is a power control coefficient, and } e_k K_i \text{ is the } k\text{-th column of } I_{K_i}. \text{ The transmitted }
\]

\(^2\) A UE that completely receives data from the BS will become inactive.

\(^3\) One can let only the participating UEs send their pilots in session \( i \), i.e., \( \tau_p = K_i \) to increase the small-scale coherence block length for payload data transmission. However, since \( \tau_c \) is normally much larger than \( K \geq K_i \) in many applications [3], letting all the UEs send their pilots, i.e., taking \( \tau_p = K \), has a negligible impact on data rates. On the other hand, the channel estimation is better when the pilot length \( \tau_p = K > K_i \), which potentially improves the data rates.

![Fig. 1. One transmission period with three UEs.](image-url)

**Remark 1.** In this work, in order to focus on fundamental principles of the proposed scheme, that is, UE assignment and rate allocation, we consider independent Rayleigh channels. The optimization and analysis of a session-based scheme tailored to correlated channels is interesting, but analytically challenging and beyond the scope of the paper. Thus, such designs are left for future work.

### IV. COMPLETION TIME MINIMIZATION

We address the minimization of the completion time of our proposed session-based scheme, and specifically minimizing the longest completion time among the UEs. For comparison, we also include the completion time minimization for the conventional schemes.

**A. Proposed Session-Based Scheme**

The problem of minimizing the completion time of UEs by optimizing the UE assignment \((a)\) and transmit power \((\eta)\) in the session-based design is

\[
\min \max_k \sum_{i \in K} a_{k,i} t_i, \tag{10a}
\]
Find a globally optimal solution to problem (10) is challenging due to the mixed-integer and nonconvex constraints (1), (4), and (7). Thus, we instead propose an approach that is suitable for practical implementation. First, we replace constraint (4) by
\[ \eta_{k,i} \leq a_{k,i} \forall k,i, \tag{11} \]
and constraint (7) by
\[ S_{k,i} \geq R_{k,i}(\eta_{k,i}) \forall k,i, \tag{12} \]
\[ S_{k,i} \geq R_{k,i}(\eta_{k,i}) \forall k,i. \tag{13} \]
We further replace constraints (12) and (13) by
\[ \tilde{\ell}_{k,i} \leq R_{k,i}(\eta_{k,i}) \forall k,i, \tag{14} \]
\[ \tilde{\ell}_{k,i} \geq a_{k,i} \forall k,i, \tag{15} \]
\[ \tilde{\ell}_{k,i} \geq a_{k,i} \forall k,i, \tag{16} \]
\[ S_{k,i} \leq \tilde{r}_{k,i} \tilde{t}_{k,i} \forall k,i, \tag{17} \]
\[ S_{k,i} \leq \tilde{r}_{k,i} \tilde{t}_{k,i} \forall k,i, \tag{18} \]
\[ \text{where } \tilde{r} \triangleq \{\tilde{r}_{k,i}\}, \tilde{r} \triangleq \{\tilde{r}_{k,i}\}, \tilde{t} \triangleq \{\tilde{t}_{k,i}\}, \tilde{t} \triangleq \{\tilde{t}_{k,i}\} \text{ are additional variables. We observe from (14)–(18) that } S_{k,i} \leq \tilde{r}_{k,i} \tilde{t}_{k,i} \forall k,i. \text{ Thus, (19) is equivalent to} \]
\[ V_{1}(\tilde{r}, \tilde{t}, S) \triangleq \sum_{k \in K, i \in K} (\tilde{r}_{k,i} \tilde{t}_{k,i} - S_{k,i}) \leq 0. \tag{20} \]
Now, in order to handle the binary constraint (1), we note that \( x \in \{0, 1\} \Leftrightarrow x \in [0, 1] \& x - x^2 \leq 0 \). Thus, (1) can be replaced by the following equivalent constraints:
\[ v_{2}(a) \triangleq \sum_{k \in K, i \in K} (a_{k,i} - a_{k,i}^2) \leq 0 \tag{21} \]
\[ 0 \leq a_{k,i}, 1 \forall k,i. \tag{22} \]
Then, problem (10) is written into a more tractable form as
\[ \min_{\tilde{a}} q \tag{23a} \]
\[ \text{s.t. (2), (3), (5), (8), (9), (10b), (11), (14)–(18), (20)–(22) } \tag{23b} \]
where \( \tilde{a} \triangleq \{\tilde{x}, \tilde{r}, \tilde{r}, \tilde{t}, q\} \), and \( q \) is an additional variable. Let \( F \triangleq \{(2), (3), (5), (8), (9), (10b), (11), (14)–(18), (20)–(22), (23b)\} \) be the feasible set of problem (23). We consider the problem
\[ \min_{\tilde{a} \in F} L(a, r, \tilde{r}, t, S, \lambda), \tag{24} \]
where \( L(a, \tilde{r}, \tilde{t}, \tilde{S}, \lambda) \triangleq q + \lambda (\gamma_{1} V_{1}(\tilde{r}, \tilde{t}, \tilde{S}) + \gamma_{2} V_{2}(a)) \) is the Lagrangian of (23), \( \gamma_{1}, \gamma_{2} > 0 \) are fixed weights, and \( \lambda \) is the Lagrangian multiplier corresponding to constraints (20), (21).
Here, \( F \triangleq F \setminus \{(20), (21)\} \).

**Proposition 1.** The values \( V_{1,\lambda} \) and \( V_{2,\lambda} \) of \( V_{1} \) and \( V_{2} \) at the solution of (24) corresponding to \( \lambda \) converge to 0 as \( \lambda \rightarrow +\infty \). Moreover, problem (23) has strong duality, i.e.,
\[ \min_{\tilde{a} \in F} q = \sup_{\lambda \geq 0} \min_{\tilde{a} \in F} L(a, \tilde{r}, \tilde{t}, \tilde{S}, \lambda), \tag{25} \]
and consequently, (23) is equivalent to (24) as the optimal solution \( \lambda^{*} \geq 0 \) of the sup-min problem in (25).

**Proof.** See Appendix. 

Theoretically, it is required to have \( V_{1,\lambda} = 0 \) and \( V_{2,\lambda} = 0 \) in order to obtain the optimal solution to problem (23). By Proposition 1, \( V_{1,\lambda} \) and \( V_{2,\lambda} \) converge to 0 as \( \lambda \rightarrow +\infty \). In practice, it is sufficient to accept \( V_{1,\lambda} \leq \varepsilon, V_{2,\lambda} \leq \varepsilon \) for some small \( \varepsilon \) with a sufficiently large value of \( \lambda \). In our numerical experiments, for \( \varepsilon = 10^{-5} \), we see that \( \lambda = 1 \) with \( \gamma_{1} = 0.1, \gamma_{2} = 0.01 \) is enough to ensure that \( V_{1,\lambda} \leq \varepsilon, V_{2,\lambda} \leq \varepsilon \). This way of choosing \( \lambda \) has been widely used in the literature, e.g., see [9] and references therein.

Problem (24) is still difficult to solve due to the nonconvex constraints (14)–(18), and nonconvex parts \( V_{1}(a), V_{2}(\tilde{r}, \tilde{t}, \tilde{S}) \) in the cost function \( L(a, r, \tilde{r}, t, S, \lambda) \). To deal with (14), we observe that \( (1 + \frac{x}{y}) \geq 1 + \frac{y^{(n)}}{y^{(n)}} + \frac{2x^{(n)}}{y^{(n)} + y^{(n)}} \).

Therefore, the concave lower bound \( \tilde{R}_{d,k,i}(\eta_{k,i}) \) of \( R_{d,k,i}(\eta_{k,i}) \) is given by \( \tilde{R}_{d,k,i} \triangleq \frac{\tau_{r,S} \log_{2}(1 + \frac{y^{(n)}}{y^{(n)}} + \frac{2y^{(n)}}{y^{(n)} + y^{(n)}})}{\tau_{r,S} \log_{2}(1 + \frac{y^{(n)}}{y^{(n)}}} - \frac{(x^{(n)} - y^{(n)}) + (x^{(n)} - y^{(n)})^{2}}{(x^{(n)} + y^{(n)})^{2}} \).

Thus, constraint (15) can be approximated by the following convex constraint
\[ \tilde{r}_{k,i} \leq \tilde{R}_{d,k,i}(\eta_{k,i}) \forall k,i. \tag{26} \]
To deal with constraints (15), we observe that \( (1 + \frac{x}{y}) \leq \log_{2} \left( \frac{x^{(n)} + y^{(n)}}{x^{(n)} + y^{(n)}} \right) \). Therefore, the convex upper bound \( \tilde{R}_{d,k,i}(\eta_{k,i}) \) of \( R_{d,k,i}(\eta_{k,i}) \) is expressed as \( \tilde{R}_{d,k,i} \triangleq \frac{\tau_{r,S} \log_{2}(1 + \frac{y^{(n)}}{y^{(n)}} + \frac{2y^{(n)}}{y^{(n)} + y^{(n)}})}{\tau_{r,S} \log_{2}(1 + \frac{y^{(n)}}{y^{(n)}})} + \frac{\frac{y^{(n)}}{y^{(n)} + y^{(n)}}}{\frac{y^{(n)}}{y^{(n)} + y^{(n)}}} - \log_{2}(\Phi_{i}) \).

Next, we observe that \( xy \leq 0.25(x+y)^{2} - 2(x^{(n)} - y^{(n)})^{2} \) and \( -xy \leq 0.25(x-y)^{2} - 2(x^{(n)} - y^{(n)})^{2} \).

Therefore, (16)–(18) can be approximated by the following convex constraints \( \tilde{d}_{k,i} \leq 0.25[(a_{k,i} - t^{(n)}_{d,k,i}) - 2(a_{k,i} + t^{(n)}_{d,k,i})](a_{k,i} + t^{(n)}_{d,k,i}) \) and \( (a_{k,i} - t^{(n)}_{d,k,i})^{2} \leq 0, \forall k,i \).

Similarly, the convex upper bounds \( V_{1}(a), V_{2}(\tilde{r}, \tilde{t}, \tilde{S}) \) of the nonconvex parts \( V_{1}(a), V_{2}(\tilde{r}, \tilde{t}, \tilde{S}) \) are respectively given by
\[ V_{1}(\tilde{r}, \tilde{t}, \tilde{S}) \triangleq \sum_{i \in K, k \in K} 0.25[(\tilde{r}_{d,k,i} + \tilde{t}_{d,k,i})^{2} - 2(\tilde{r}_{d,k,i} + \tilde{t}_{d,k,i})] \]
\[ V_{2}(a) \triangleq \sum_{i \in K, k \in K} (a_{k,i} - 2a_{k,i}^{(n)} a_{k,i} + a_{k,i}^{(n)}). \tag{30} \]
At iteration \( n+1 \), for a given point \( \tilde{x}^{(n)} \), problem (24) can finally be approximated by the following convex problem
\[ \min_{\tilde{a} \in F} L(a, \tilde{r}, t, S, \lambda), \tag{31} \]
where \( \bar{L}(a, \beta, i, S, \lambda) \triangleq q + \lambda(\gamma_1 V_1(a) + \gamma_2 V_2(\bar{r}, i, S)) \) and \( \bar{F} \triangleq \{ (2), (3), (5), (10b), (11), (22), (23b), (26), (27) - (30) \} \) is a convex feasible set. In Alg. 1, we outline the main steps to solve problem (24). Starting from a random point \( \bar{x} \in \bar{F} \), we solve (31) to obtain its optimal solution \( \bar{x}^* \), and use \( \bar{x}^* \) as an initial point in the next iteration. The algorithm terminates when an accuracy level of \( \varepsilon \) is reached. Alg. 1 will converge to a Fritz John solution of problem (24) (hence (23) or (10)). The proof of this fact is rather standard, and follows from [9, Proposition 2].

Note that Problem (10) is constructed using the achievable rates (5) that depend only on the large-scale coefficients \( \beta \). Before the downlink transmission, the BS solves (10) to obtain the session durations, per-session user assignments, data rates, and transmit powers. Therefore, no extra signalling overhead to schedule UEs/rates and no optimization algorithm are required during the transmission.

B. Conventional Schemes

The conventional schemes can be considered as special cases of the session-based scheme with only one session. Thus, all variables in the conventional schemes can be directly obtained from the session-based scheme by dropping the index \( i \). More precisely, the power constraint at the BS and the achievable rate of UE \( k \) in the conventional schemes are, respectively, given by

\[
\sum_{\ell \in K} \rho_{k, \ell} \leq 1
\]

and \( R_k(\eta) = \frac{1}{\log_2(1 + \text{SINR}_k(\eta))} \), where \( \eta \triangleq \{ \eta_{k, \ell} \}_{\ell \in K} \) are power control coefficients, and \( \text{SINR}_k(\eta) = \frac{1}{\rho(\beta_k - \sigma_k^2) \sum_{\ell \in K} \eta_{k, \ell}} \). Note that since \( T_c \) is in order of milliseconds, and the conventional schemes have only one session, the completion times of UEs is normally larger than \( T_c \), which is confirmed in the numerical results in Section V.

1) Conventional Data Size-Aware Scheme: The corresponding problem of completion time minimization is

\[
\min \eta \sum_{k \in K} S_k \frac{S_k}{R_k(\eta)} \quad \text{s.t.} \quad (32)
\]

\[
0 \leq \eta_{k, \ell}, \forall k,
\]

\[
S_k/R_k(\eta) \leq \tilde{T}_c,
\]

Problem (33) can be transformed into epigraph form as

\[
\min z, \quad \text{s.t.} \quad (32), (33b)
\]

\[
S_k/r_k \leq z \leq \tilde{T}_c, \forall k
\]

\[
r_k \leq R_k(\eta), \forall k
\]

where \( \eta \triangleq \{ \eta, r, z \} \), \( r \triangleq \{ r_k \}, \forall k \), and \( z \) are additional variables. Using the same approach to deal with constraint (14), we obtain the concave lower bound \( \tilde{R}_k(\eta) \) of \( R_k(\eta) \) as

\[
\tilde{R}_k \triangleq \frac{1}{\lambda_0 \log_2 B} \left[ \log \left( 1 + \frac{\gamma_k}{\sigma_k^2} \right) + \frac{2\gamma_k}{\gamma_k + \phi_k} - \frac{(\gamma_k)^2}{\gamma_k + \phi_k} \right],
\]

where \( \gamma(\eta_k) \triangleq (M - K)\rho_0^2 k, \Phi(\eta) \triangleq \rho(\beta_k - \sigma_k^2) \sum_{\ell \in K} \eta_{k, \ell} + 1 \). Then, constraints (34c) can be approximated by the following convex constraint

\[
r_k \leq \tilde{R}_k(\eta), \forall k.
\]

Now, at iteration \((n+1)\), for a given point \( \eta^{(n)} \), problem (34) can be approximated by the following convex problem:

\[
\min z, \quad \text{s.t.} \quad (32), (33b)
\]

\[
S_k/r_k \leq z \leq \tilde{T}_c, \forall k
\]

\[
r_k \leq \tilde{R}_k(\eta), \forall k
\]

Algorithm 2 Solving problem (34)

1: Initialize: Set \( n=0 \) and choose a random point \( \eta^{(0)} \in \tilde{H} \).
2: repeat
3: Update \( n = n + 1 \)
4: Solve (36) to obtain its optimal solution \( \eta^* \)
5: Update \( \eta^{(n)} = \eta^* \)
6: until convergence

where \( \tilde{H} \triangleq \{ (32), (33b), (34b) \} \) is a convex feasible set. In Alg. 2, we outline the main steps to solve problem (34). Let \( \tilde{H} \triangleq \{ (32), (33b), (34b) \} \) be the feasible set of problem (34). Starting from a random point \( \eta \in \tilde{H} \), we solve (36) to obtain its optimal solution \( \eta^* \), and use \( \eta^* \) as an initial point in the next iteration. The algorithm terminates when an accuracy level of \( \varepsilon \) is reached. Since \( \tilde{H} \) satisfies Slater’s constraint qualification condition, Alg. 2 converges to a Karush–Kuhn–Tucker solution of (34) (hence (33)) [11, Theorem 1].

2) Conventional Non-Data Size-Aware Scheme: This scheme does not take into account the size of the UE data. The completion time is reduced by improving the rates of all UEs. To this end, the scheme aims to maximize the lowest rate of all UEs, which leads to the following problem

\[
\max \eta_k \min \tilde{R}_k(\eta)
\]

\[
\text{s.t.} \quad (32), (33b)
\]

The optimal solution to problem (37) can be written in closed-form as

\[
\eta_k = \frac{\log_2(1 + \frac{1}{1 + \beta_k - \sigma_k^2} \sum_{\ell \in K} \eta_{k, \ell})}{\log_2(1 + \frac{1}{1 + \beta_k - \sigma_k^2} \sum_{\ell \in K} \eta_{k, \ell})}
\]

3) Heuristic Small-Scale-Fading-Based Scheme: In each small-scale coherence block, greedy user scheduling and power allocation to (approximately) maximize the lowest rate are performed. Specifically, if the data queue of a UE becomes zero, this UE will be no longer scheduled in the subsequent small-scale coherence block. This way, the remaining UEs at later small-scale coherence blocks will have more power and higher transmission rate, which eventually contributes to reducing the longest completion time. The optimal power control for the max-min rate problem is

\[
\eta_k = \frac{1}{c_k / \left( \prod_{k \in K} (1/c_k) \right)},
\]

where \( c_k = \beta_k g_k^2 \).
the completion time compared with conventional schemes. By the definition of $L_{\lambda_1}$ and $L_{\lambda_2}$,

$$L_{\lambda_1} = q_{\lambda_1} + \lambda_1 V_{\lambda_1} \leq q_{\lambda_2} + \lambda_2 V_{\lambda_2},$$

(39)

and

$$L_{\lambda_2} = q_{\lambda_2} + \lambda_2 V_{\lambda_2} \leq q_{\lambda_1} + \lambda_1 V_{\lambda_1} + \lambda_2 V_{\lambda_2},$$

(40)

from which we have $\lambda_1 V_{\lambda_1} + \lambda_2 V_{\lambda_2} \leq \lambda_1 V_{\lambda_2} + \lambda_2 V_{\lambda_1}$, and so $V_{\lambda_2} \leq V_{\lambda_1}$. This means $V_{\lambda}$ is decreasing as $\lambda$ is increasing. Since $V_{\lambda} = \gamma_1 V_{\lambda_1} + \gamma_2 V_{\lambda_2} \geq 0$ for all $\lambda \geq 0$, we obtain that $V_{\lambda} \rightarrow V^* \geq 0$ as $\lambda \rightarrow +\infty$. From (39) and (40), we also have that $\lambda q_{\lambda_2} + \lambda_1 q_{\lambda_2} \leq \lambda q_{\lambda_2} + \lambda_1 q_{\lambda_1}$, which yields $q_{\lambda_2} \geq q_{\lambda_1}$. Therefore, as $\lambda \rightarrow +\infty$, $q_{\lambda}$ is increasing and hence bounded from below. Now, if $V^* > 0$, then $L_{\lambda} = q_{\lambda} + \lambda V_{\lambda} \rightarrow +\infty$ as $\lambda \rightarrow +\infty$, which contradicts (38). Thus, we must have $V^* = 0$, that is, $V_{\lambda} \rightarrow 0$ as $\lambda \rightarrow +\infty$, which implies that $V_{1,\lambda} \rightarrow 0$ and $V_{2,\lambda} \rightarrow 0$ as $\lambda \rightarrow +\infty$.

Finally, let $\tilde{\lambda}$ be the value of $\lambda$ corresponding to $\lambda$. Then $\tilde{\lambda} \in \mathcal{F}$. Since $\mathcal{F}$ is bounded, there exists a cluster point $\tilde{\lambda}$ of $\{\tilde{\lambda}_{\alpha}\}$ as $\lambda \rightarrow +\infty$. We assume, without loss of generality, that $\tilde{\lambda} \rightarrow \lambda_*$. Then, \( a_{\lambda} \to a_* \), \( \tilde{r}_{\lambda} \to \tilde{r}_* \), \( S_{\lambda} \to S_* \), and $t_{\lambda} \to t_*$. It follows that $V_{1,\lambda} \rightarrow (V_1)_{\#} = \sum_{k \in K} \sum_{i \in K} (\tilde{r}_{k,i} - (S_{k,i})_{\#})$, $V_{2,\lambda} \rightarrow (V_2)_{\#} = \sum_{k \in \mathcal{N}} \sum_{i \in K} ((a_{k,i})_{\#} - (a_{k,i})^2_{\#})$, $V_{\lambda} \rightarrow V_{\#} = \gamma_1 (V_1)_{\#} + \gamma_2 (V_2)_{\#}$, and $q_{\lambda} \rightarrow q_*$. As above, $(V_1)_{\#} = 0$, $(V_2)_{\#} = 0$, and $V_{\#} = 0$. Therefore, $(\tilde{r}_*, t_*, S_*)$ and $(a_*)$ satisfy (20) and (21), respectively. This together with $\tilde{\lambda} \in \mathcal{F}$ implies that $\tilde{\lambda} \in \mathcal{F}$, and so $\tilde{\lambda}$ is a feasible point of (23). As such, $q_* \geq q^*$. Next, the definition of $\tilde{\lambda}$ implies that, for all $\lambda \geq 0$, $\sup_{\lambda \geq 0} L_{\lambda} \geq \tilde{\lambda} \geq q_0 + \lambda V_{\lambda} \geq q_0$. By letting $\lambda \rightarrow +\infty$, $\sup_{\lambda \geq 0} L_{\lambda} \geq q_* + \lambda V_{\lambda} \geq q_*$. Combining with (38), yields $\sup_{\lambda \geq 0} \tilde{L}(\lambda) = q_* = q^*$. We conclude that (25) holds and that $\tilde{\lambda}$ is an optimal solution of (23), which completes the proof.

REFERENCES