Dynamic Scheduling with Statistical Delay Guarantees and Traffic Dropping

Khoa T. Phan, and Tho Le-Ngoc
Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
Email: khoa.phan@mail.mcgill.ca, tho.le-ngoc@mcgill.ca

***Abstract*—This work studies the dynamic scheduling problems in wireless networks with delay-sensitive loss-tolerant users. The users' traffic satisfies some statistical delay constraints. Moreover, the traffic can be dropped but the dropping rates do not exceed some thresholds. We consider two scheduling scenarios. First, we study the problem to minimize the total transmission power while maintaining the minimum rates for the users. Then, we study the problem to maximize the minimum rate(s) of the users while constraining the maximum total power. We derive the optimal solutions for both scheduling problems. When the fading statistics are available, using the dual-gradient method, the optimal policies can be computed. When the fading statistics are unknown, this work proposes online scheduling algorithms using online time-averaging. The convergence and optimality of the proposed algorithm are guaranteed by the results in stochastic approximation theory.**

–Key words–: Dynamic scheduling, effective capacity, dropping rate, quality of service (QoS), stochastic approximation.

I. Introduction

In wireless communications, dynamic (or opportunistic) scheduling exploiting the multi-user fading diversity and temporal fading diversity is an effective mechanism to manage interference and hence, improving the spectrum utilization [1]. While there have been many works on dynamic scheduling under different network settings, the ones more related to our current work are [2], [3], [4]. The work in [2], [3] proposes scheduling policies with minimum rate constraints for the users. The traffic is assumed to be delay insensitive and only one user is scheduled in each slot. The scheduling problems can be solved for the optimal solutions. Taking traffic delay into consideration, in [4], the power optimal scheduling policies with and without traffic dropping for guaranteed minimum rate and statistical delay for each user are developed. Users are assumed to time-share the channel (i.e., each user transmits for a non-overlapping fraction of time in each slot),[1] and the scheduling problem is convex and can be solved for optimal solutions. In [2]– [4], online scheduling algorithms are proposed without requiring known fading statistics.

This current work considers dynamic scheduling problems for delay-sensitive loss-tolerant users (or applications), for example, voice over IP or multimedia transmission over wireless networks. Users are allowed to drop some of their traffic but the traffic dropping rates cannot exceed some thresholds. Moreover, the users' traffic should satisfy some statistical delay constraints. Different from [4], users are assumed to transmit one at a time. We study two different scheduling formulations. First, we study the problem to minimize the total transmission power while maintaining the minimum rates with statistical delay guarantees for the users. Then we study the problem to maximize the minimum rate(s) of the users while constraining the maximum total power. We derive the optimal solutions for both scheduling problems. When the fading statistics are available, using the dual-gradient method, the optimal policies can be computed. Moreover, when the fading statistics are unknown, this work proposes online scheduling algorithms using online time-averaging. Illustrative results demonstrate the performance of the proposed scheduling algorithms in various settings.

[1] In terms of implementation, time-sharing scheduling is usually applicable for downlink communications only while exclusive channel assignment scheduling can be used for both downlink and uplink communications.

II. Problem Formulations

A. System model and scheduling metrics

Consider a single channel multi-user system with a centralized scheduler. Traffic for each user is stored in its own First-In-First-Out (FIFO) buffer. It is assumed that there is always a sufficient amount of backlogged data for scheduling purposes. Time is divided into slots of equal duration.

The channel state representing the power gain in slot t is denoted by the vector $\mathbf{h}^t = (h_1^t, h_2^t, \ldots, h_N^t) \in \mathcal{H}^N$, where $h_i^t \in \mathcal{H}$ denotes the channel state of user i, $\mathcal{H}$ denotes the set of possible states, and N is the number of users. Let $\mathcal{N} = \{1, \ldots, N\}$ be the (index) set of all users. The channel state is assumed to be time-varying block-fading over the time slots. We assume that $\{h_i^t\}, \forall i, t$ are independent and identically distributed (i.i.d.) over $\mathcal{H}$ with some general probability distribution function (PDF) ν. In slot t, for user i, $i \in \mathcal{N}$, we associate an indicator $y_i^t \in \{0, 1\}$ which is 1 if user i is scheduled, otherwise 0. Since only one user is active in each slot, we have $\sum_{i \in \mathcal{N}} y_i^t = 1, \forall t$. If user i is scheduled in slot t, let $z_i^t \geq 0$ and $p_i^t \geq 0$ denote the amount of dropped traffic from user i's buffer and the transmission power, respectively. Hence, $z_k^t = p_k^t = 0$ for all non-scheduled users $k \neq i$ in slot t. The information theoretic rate r with power p under channel state h in a slot is considered:

$$r(p, h) = \log(1 + ph).$$

The total power is defined:

$$\mathcal{P} \triangleq \lim_{T\to\infty} \frac{1}{T}\left\{\sum_{t=1}^{T}\sum_{i=1}^{N} p_i^t y_i^t\right\} = \mathbb{E}\left\{\sum_{i=1}^{N} p_i y_i\right\} \quad (1)$$

where $\mathbb{E}\{.\}$ denotes the statistical expectation with respect to (w.r.t.) the fading distribution ν. The second equality holds due to ergodicity.

Each user's traffic satisfies a statistical delay QoS constraint. In particular, it is required that the probability for the queue length of user i exceeding a certain threshold, x, decays exponentially as a function of x. Such exponential decay is characterized by a delay exponent θ_i [5]:

$$\theta_i \triangleq -\lim_{x\to\infty} \frac{\log\big(\Pr\{q_i(\infty) > x\}\big)}{x} \in [0,\infty) \quad (2)$$

where $q_i(\infty)$ denotes user i's buffer length at equilibrium and $\Pr(a > b)$ is the probability that the inequality $a > b$ holds. A higher θ_i means more stringent delay requirements for user i and vice versa. Under the i.i.d. block fading channels assumption, the effective capacity of user i with delay exponent θ_i is given by [5]:

$$\mathcal{C}_i(\theta_i) \triangleq -\frac{1}{\theta_i}\log\left(\mathbb{E}\Big\{e^{-\theta_i y_i(z_i + r(p_i,h_i))}\Big\}\right). \quad (3)$$

$\mathcal{C}_i(\theta_i)$ is interpreted as user i's maximum constant arrival rate that the scheduling algorithm can support in order to guarantee the delay requirement specified by θ_i.

The dropping rate of user i is defined as the fraction between the amount of dropped traffic over the total amount of traffic leaving the buffer:

$$\mathcal{L}_i \triangleq \lim_{T\to\infty} \frac{\sum_{t=1}^{T} y_i^t z_i^t}{\sum_{t=1}^{T} y_i^t (r(p_i^t, h_i^t) + z_i^t)} = \frac{\mathbb{E}\{y_i z_i\}}{\mathbb{E}\{y_i(r(p_i,h_i)+z_i)\}}.$$

B. Dynamic scheduling problem formulations

1) Power-minimization scheduling: We study the scheduling problem to minimize the power while maintaining the minimum effective capacities and the maximum dropping rates for the users. In each slot, the scheduling policy determines which user should transmit and what should be its transmission power and the amount of traffic dropped from its buffer. Hence, the problem can be formally posed as:

$$\begin{aligned} \text{minimize} \quad & \mathcal{P} \\ \text{such that:} \quad & \mathcal{C}_i(\theta_i) \geq C_i^{\text{ef}}, \quad \forall i \in \mathcal{N} \\ & \mathcal{L}_i \leq L_i, \quad \forall i \in \mathcal{N} \end{aligned} \quad (4)$$

where C_i^{ef} and $L_i \in [0,1)$ are, respectively, the minimum effective capacity and maximum dropping rate requirements of user i. The traffic loss requirement gets more stringent as L_i decreases and no traffic is dropped when $L_i = 0$.

2) Min-rate maximization scheduling: In the second problem, we aim at maximizing the minimum rate(s) of the users under the maximum power constraint. The scheduling problem can be stated as:

$$\begin{aligned} \text{maximize} \quad & \min_{i\in\mathcal{N}} \Big\{\mathcal{C}_i(\theta_i)\Big\} \\ \text{such that:} \quad & \mathcal{P} \leq P^{\max} \\ & \mathcal{L}_i \leq L_i, \quad \forall i \in \mathcal{N} \end{aligned} \quad (5)$$

where $P^{\max}$ is the maximum power constraint.

III. Power-Minimization Scheduling

A. Optimal policy characterization

The optimal solution of (4) is given in Theorem 1.[2]

Theorem 1: For channel state $\mathbf{h} = (h_1, \ldots, h_N)$, the optimal policy is to schedule user k with power p_k and amount of dropped traffic z_k where:

$$k = \arg\min_{i\in\mathcal{N}} \Big\{\mathcal{F}_i\Big\} \quad (6)$$

and the scheduling parameters for user $i \in \mathcal{N}$ are given by:

$$\begin{aligned} \mathcal{F}_i = {} & p_i + \lambda_i e^{-\theta_i(z_i + \log(1+p_i h_i))} \\ & + \beta_i\Big((1-L_i)z_i - L_i\log(1+p_i h_i)\Big) \end{aligned} \quad (7)$$

$$p_i = \Big[\beta_i - \frac{1}{h_i}\Big]^+ \quad (8)$$

$$z_i = \left[-\frac{1}{\theta_i}\log\left(\frac{\beta_i(1-L_i)}{\lambda_i\theta_i\Big(1+\big[\beta_i - \frac{1}{h_i}\big]^+ h_i\Big)^{-\theta_i}}\right)\right]^+ \quad (9)$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$ are the multipliers of (4) associated with the capacity and dropping rate constraints, respectively; $[x]^+$ denotes $\max\{0, x\}$ for some number x.

At optimality, the inequality constraints are met with equalities and the multipliers $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$ are strictly positive.

The scheduling policy takes into account both the channel conditions and QoS requirements of all the users. The power allocation (8) is water-filling solution with users having different water levels. The amount of dropped traffic z_k for the scheduled user k depends on the user k's channel conditions. For favorable channel conditions, more traffic is transmitted and less traffic is dropped for the scheduled user and vice versa.

We say that the users are homogeneous when $\theta_i = \theta$, $C_i^{\text{ef}} = C^{\text{ef}}$, and $L_i = L, \forall i \in \mathcal{N}$.

Lemma 1: For homogeneous users, the optimal policy for (4) selects the user with largest channel state in each slot.

In general, the multipliers $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$ can be found using dual-gradient iterations for user $i \in \mathcal{N}$:

$$\lambda_i^{t+1} = \lambda_i^t + \epsilon\left(\mathbb{E}\Big\{e^{-\theta_i y_i\left(z_i + \log(1+p_i t h_i)\right)}\Big\} - e^{-\theta_i C_i^{\text{ef}}}\right) \quad (10)$$

$$\beta_i^{t+1} = \beta_i^t + \epsilon\mathbb{E}\Big\{y_i\big((1-L_i)z_i - L_i\log(1+p_i h_i)\big)\Big\} \quad (11)$$

[2]Due to space limitation, we omit the proofs for the presented results.

where ϵ is some small positive coefficient and the scheduling solutions y_i, z_i, and p_i for the channel state $\mathbf{h}$ are given in Theorem 1 using current values $\boldsymbol{\lambda}^t$, $\boldsymbol{\beta}^t$. The iterations converge to the optimal $\boldsymbol{\lambda}$, $\boldsymbol{\beta}$ in Theorem 1 from any positive initial $\boldsymbol{\lambda}^1, \boldsymbol{\beta}^1$ [6].

B. Online power-minimization scheduling algorithm

The knowledge of the fading PDF ν is required to evaluate the expectations involved in (10)–(11). However, the PDF ν is usually unknown in real-life communications. Even when the PDF is available, it is often impossible to compute the expectations in closed-form. Hence, computing the multipliers using (10)–(11) is practically impossible. This motivates us to study online scheduling algorithm without requiring known PDF ν.

We now propose a stochastic approximation based online scheduling algorithm using similar approach as in [3]. The idea is to replace the ensemble iterations (10)–(11) with stochastic approximation iterations. Consequently, the stochastic dual-gradient iterations in slots $t = 1, 2, \ldots$ are given by:

$$\bar{\lambda}_i^{t+1} = \bar{\lambda}_i^t + \epsilon^t \Big(e^{-\theta_i y_i^t \left(z_i^t + \log(1+p_i^t h_i^t) \right)} - e^{-\theta_i C_i^{\text{ef}}} \Big) \quad (12)$$

$$\bar{\beta}_i^{t+1} = \bar{\beta}_i^t + \epsilon^t y_i^t \Big((1-L_i) z_i^t - L_i \log(1+p_i^t h_i^t) \Big) \quad (13)$$

for all users $i \in \mathcal{N}$ with positive initial values. The scheduling solutions y_i^t, z_i^t and p_i^t for user i are given by Theorem 1 using current estimates $\bar{\boldsymbol{\lambda}}^t$, $\bar{\boldsymbol{\beta}}^t$. $\{\epsilon^t\}_{t=1,2,\ldots}$ is a positive scalar learning sequence satisfying the following condition: $\sum_{t=1}^{\infty} \epsilon^t = \infty$; $\sum_{t=1}^{\infty} (\epsilon^t)^2 < \infty$.

The iterations (12)–(13) involve stochastic estimates of their counterparts (10)–(11) and are based on instantaneous (instead of average) channel and scheduling decisions. Different from (10)–(11), the multipliers $\bar{\boldsymbol{\lambda}}^t$, $\bar{\boldsymbol{\beta}}^t$ are updated online in each slot without requiring known fading PDF. The convergence of the proposed algorithm (12)–(13) is provided next.

Proposition 1: We have $\lim_{t\to\infty} \bar{\lambda}_i^t = \lambda_i$ and $\lim_{t\to\infty} \bar{\beta}_i^t = \beta_i$ for all $i \in \mathcal{N}$.

The proposed online learning algorithm does not assume any specification on the fading statistics and converges for any i.i.d. fading distribution. Hence, it is very robust to channel model variations. Moreover, the speed of convergence depends on the step sizes and the initial multiplier values. However, such convergence study is out of the scope of this work.

IV. Min-rate Maximization Scheduling

A. Optimal policy characterization

Using a new variable C, the scheduling problem (5) can be recast as:

$$\begin{aligned} \text{maximize} \quad & C \\ \text{such that:} \quad & \mathcal{C}_i(\theta_i) \geq C, \quad \forall i \in \mathcal{N} \\ & \mathcal{P} \leq P^{\max} \\ & \mathcal{L}_i \leq L_i, \quad \forall i \in \mathcal{N} \end{aligned} \quad (14)$$

The optimal solution of (14) is given as follows.

Theorem 2: For channel state $\mathbf{h} = (h_1, \ldots, h_N)$, the optimal policy is to schedule user k with power p_k and the amount of dropped traffic z_k where:

$$k = \arg\min_{i\in\mathcal{N}} \left\{ \bar{\mathcal{F}}_i \right\} \quad (15)$$

and the scheduling parameters for user $i \in \mathcal{N}$ are given by:

$$\begin{aligned} \bar{\mathcal{F}}_i &= \kappa p_i + \lambda_i e^{-\theta_i (z_i + \log(1+p_i h_i))} \\ &\quad + \beta_i \Big((1-L_i) z_i - L_i \log(1+p_i h_i) \Big) \quad (16) \\ p_i &= \Big[\frac{\beta_i}{\kappa} - \frac{1}{h_i} \Big]^+ \quad (17) \\ z_i &= \left[-\frac{1}{\theta_i} \log \left(\frac{\beta_i (1-L_i)}{\lambda_i \theta_i \Big(1 + \big[\frac{\beta_i}{\kappa} - \frac{1}{h_i} \big]^+ h_i \Big)^{-\theta_i}} \right) \right]^+ \quad (18) \end{aligned}$$

where κ, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)$ are the multipliers of (14) associated with the power, capacity and loss rate constraints, respectively. The optimal min-rate effective capacity C^* is given by:

$$\sum_{i=1}^{N} \lambda_i \theta_i e^{-\theta_i C^*} = 1.$$

Again, at optimality, the inequalities are met with equalities and the Lagrange multipliers are positive.

Lemma 2: For homogeneous users, the optimal policy for (14) selects the user with largest channel state in each slot.

Similar to the power minimization scheduling problem, the multipliers in Theorem 2 can be computed using dual-gradient iterations when the fading PDF ν is known. We omit the details here.

B. Online min-rate maximization scheduling algorithm

When the PDF ν is unknown, the following online scheduling algorithm can be used to compute the multipliers in Theorem 2. The stochastic dual-gradient iterations for each user $i \in \mathcal{N}$ in slot $t = 1, 2, \ldots$ are:

$$\bar{\lambda}_i^{t+1} = \bar{\lambda}_i^t + \epsilon^t \Big(e^{-\theta_i y_i^t (z_i^t + \log(1+p_i^t h_i^t))} - e^{-\theta_i C^t} \Big) \quad (19)$$

$$\bar{\beta}_i^{t+1} = \bar{\beta}_i^t + \epsilon^t y_i^t \Big((1-L_i) z_i^t - L_i \log(1+p_i^t h_i^t) \Big) \quad (20)$$

$$\bar{\kappa}^{t+1} = \bar{\kappa}^t + \epsilon^t \big(p^t - P^{\max} \big) \quad (21)$$

with some initial positive values κ^1, $\boldsymbol{\lambda}^1$ and $\boldsymbol{\beta}^1$. In (21), p^t equals to the allocated power for the scheduled user in slot t. Also, C^t is computed such that:

$$\sum_{i=1}^{N} \bar{\lambda}_i^t \theta_i e^{-\theta_i C^t} = 1. \quad (22)$$

The optimal scheduling solutions z_i^t, y_i^t, and p_i^t are given in Theorem 2 using the current estimates of the multipliers. The convergence of the proposed algorithm is presented next.

Proposition 2: We have $\lim_{t\to\infty} \bar{\lambda}_i^t = \lambda_i$ and $\lim_{t\to\infty} \bar{\beta}_i^t = \beta_i$ for all $i \in \mathcal{N}$, $\lim_{t\to\infty} \bar{\kappa}^t = \kappa$, and $\lim_{t\to\infty} C^t = C^*$.

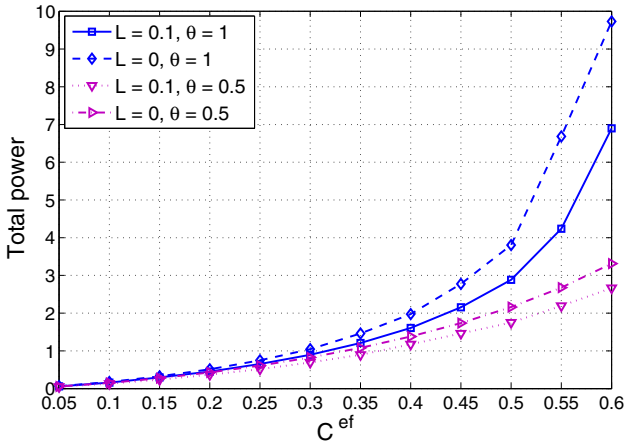

Fig. 1. Total power consumption versus capacity requirement C^{ef}.

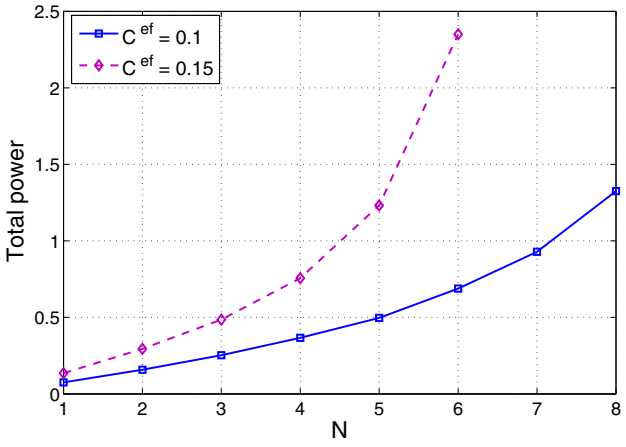

Fig. 2. Total power consumption versus number of users N.

V. Illustrative Results

A. Simulation setup

Due to lack of space, we illustrate the results of the power minimization scheduling only. Without loss of generality, users are assumed to be homogeneous. The dropping rate threshold is fixed $L = 0.1$.

As an illustrative example, the channel state space is assumed to have 8 states $\mathcal{H} = \{0.0131, 0.0418, 0.0753, 0.1157, 0.1661, 0.2343, 0.3407, 0.6200\}$ with the corresponding probabilities $[1, 1, 2, 3, 3, 2, 1, 1]/14$.

The learning duration, the learning rate sequence ϵ^t, and the initial multipliers are varied for different simulation examples to ensure convergence within the running duration.

B. Illustrative results

We first plot the power consumption versus the capacity requirements C^{ef} when $N = 2$ users for delay exponents $\theta = 1$ (more stringent delay constraint) and $\theta = 0.5$ in Fig. 1. Also, the power consumption when dropping is not allowed is also plotted, i.e., $L = 0$. It is clear that when the capacity requirement increases, more power is needed. Moreover, as the required capacity becomes larger, one needs to pay more in terms of excess power to provide the users an extra rate 0.05. By allowing traffic dropping at rate 0.1, power can be saved, especially at larger capacity requirement. For example, when $C^{\text{ef}} = 0.6$ and $\theta = 1$, traffic dropping can save approximately 30% in power consumption. Traffic dropping can save more power when the delay constraint is more stringent. It implies that more delay-sensitive applications benefit most from traffic dropping. Finally, less power is required for looser delay constraint since temporal diversity can be capitalized.

Fig. 2 shows the power consumption versus the number of users N for $C^{\text{ef}} = 0.1$ and 0.15 and $\theta = 1$. We can see that serving one more user requires more excess power when there are already a large number of users in the system. Also, more power is required for larger capacity requirements, especially when there are more users in the system.

We next examine the convergence of the learning algorithm (12)–(13). Let the parameters be $N = 2$, $L = 0.1$, $C^{\text{ef}} = 0.2$, and $\theta = 1$. The initial values for the multipliers are $\boldsymbol{\lambda}^1 = [1, 2]$ and $\boldsymbol{\beta}^1 = [1, 2]$. We choose different initial values on purpose to show that they will converge to the same values as expected. The learning sequence is $\epsilon^t = 1/t^{.55}$ and the learning duration is 10^5 slots. The estimated dropping rate and effective capacity for user i at slot t from the simulation traces are calculated as:

$$\bar{L}_i^t = \frac{\sum_{t=1}^t \bar{y}_i^t \bar{z}_i^t}{\sum_{t=1}^t \bar{y}_i^t (\bar{r}_i^t + \bar{z}_i^t)}, \quad \bar{C}_i^{\text{ef}\,t} = -\frac{1}{\theta} \log\left(\frac{1}{t} \sum_{t=1}^t e^{-\theta \bar{y}_i^t (\bar{z}_i^t + \bar{r}_i^t)} \right) \tag{23}$$

where $\bar{r}_i^t = \log(1 + \bar{p}_i^t h_i^t)$. At convergence, it is expected that $\lim_{t\to\infty} \bar{L}_i^t \approx [0.1, 0.1]$ and $\lim_{t\to\infty} \bar{C}_i^{\text{ef}\,t} \approx [0.2, 0.2]$.

We can see from Figures 3, 4, and 5 that the learning algorithm converges and at convergence, the dropping rates are $[0.1001, 0.0997]$ and the effective capacities are $[0.2000, 0.2002]$. Hence, the QoS requirements are satisfied with equalities. The convergent is reasonably fast for the chosen parameters. Let us examine the convergence of user 1, especially at the initial learning phase. During this learning phase, the estimated multiplier $\bar{\beta}_1$ of the user 1 is smaller (than that of user 2), hence, its allocated power is smaller. Consequently, user 1's achieved rate is below the target value. On the other hand, more traffic is dropped for user 1, and the dropping rate is above the threshold (see Fig. 4). Hence, when less data is transmitted, more data is dropped and vice versa.

VI. Conclusions

We have studied the dynamic scheduling problems in wireless networks with delay-sensitive loss-tolerant users. First, we studied the problem to minimize the total transmission power while maintaining the minimum rates with statistical delay

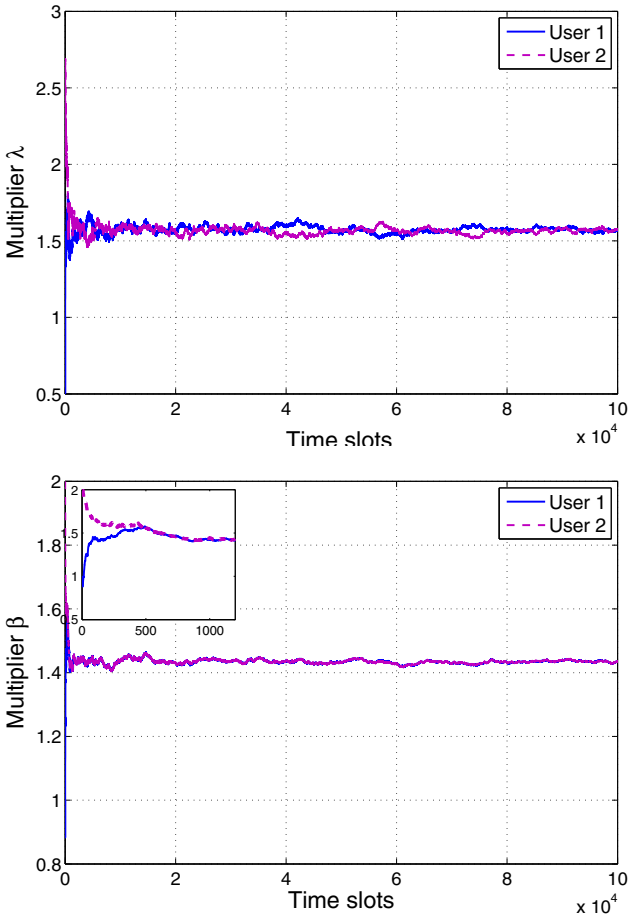

Fig. 3. Convergence of the Lagrange multipliers λ and β.

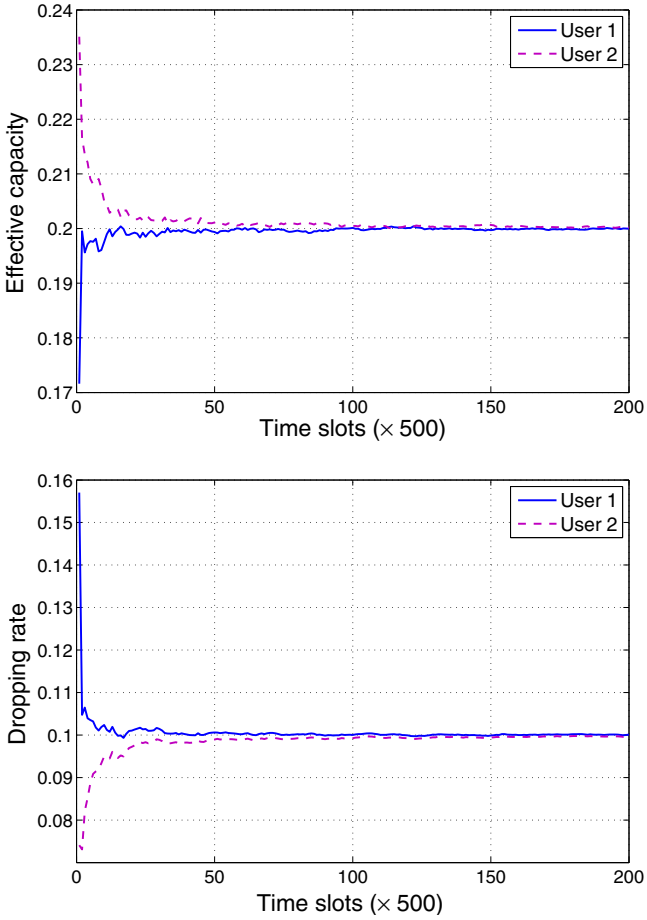

Fig. 4. Convergence of the dropping rate and effective capacity.

guarantees for the users. Then, we studied the problem to maximize the minimum rate(s) of the users while constraining the maximum total power. Optimal solutions for both scheduling problems are derived. Moreover, this work proposes online scheduling algorithms using online time-averaging without requiring a-priori known fading statistics.

REFERENCES

[1] R. Knopp, and P. A. Humblet, "Information Capacity and Power Control in Single-cell Multiuser Communications," Proc. *IEEE Int. Conf. Communs. (ICC)*, Seatle, WA, USA, June 1995.

[2] X. Liu, E. Chong, and N. Shroff, "Joint Scheduling and Power-Allocation for Interference Management in Wireless Networks," Proc. *IEEE VTC*, Vancouver, Canada, Fall 2002.

[3] A. Bhorkar, A. Karandikar, and V. S. Borkar, "Power Optimal Opportunistic Scheduling," Proc. *IEEE GLOBECOM*, San Francisco, CA, USA, Nov. 2006.

[4] Q. Du, and X. Zhang, "Statistical QoS Provisionings for Wireless Unicast/Multicast of Multi-Layer Video Streams," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 3, pp. 420–433, April 2010.

[5] D. Wu, and R. Negi, "Effective Capacity: A Wireless Link Model for Support of Quality of Service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[6] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

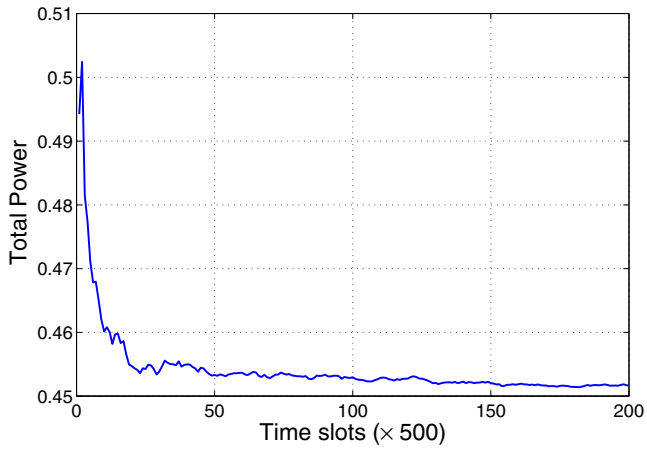

Fig. 5. Convergence of the total power consumption.