Optimal Resource Allocation for Multimedia Applications over Multiaccess Fading Channels

Cong Shen, *Student Member, IEEE,* and Mihaela van der Schaar, *Senior Member, IEEE*

***Abstract*—We study the problem of optimal resource allocation for multi-user multiaccess wireless video communication from an information-theoretic point of view. We derive the optimal resource allocation policies by directly maximizing at the application layer the weighted sum of video qualities of all users, subject to information-theoretic multiaccess capacity region constraints in the MAC-PHY layers. We solve this problem for three multiaccess capacity regions: 1) non-fading channel, 2) fading channel with a *given* power control policy, and 3) fading channel with *dynamic* power control policies. The optimal resource allocation policy is referred as Largest Quality Improvement Highest Possible Rate (LQIHPR). We propose simple greedy algorithms to implement this policy. Since the capacity region is the fundamental characterization of achievable rates, the solutions developed in this paper provide the operational upper bound of achievable video quality in a multiaccess fading channel.**

***Index Terms*—Cross-layer design and optimization, resource management, multimedia communication, multiuser channels.**

I. Introduction

RECENT research has shown that significant performance gains can be achieved by resource allocation in wireless networks. As a result of the time-varying nature of wireless channels, resource allocation in wireless networks has been extensively studied, focusing on several different aspects such as increasing throughput, minimizing delay, improving fairness, etc. In this paper, we focus on studying the resource (power and rate) allocation problem in a multi-user multimedia transmission environment. Our work approaches this problem from two unique perspectives:

1) The ultimate goal of the resource allocation for multimedia users is to maximize their video quality. This is in contrast with the majority of existing research on resource allocation, which focuses on the interaction among PHY, MAC and Network layers. The existing results for resource allocation have the advantage that they are simple and transparent to high layers. However, since they only consider the lower layers, they are not optimal from a video quality perspective.
2) The constraint of the resource allocation is the throughput capacity region defined at the PHY/MAC layers. In contrast, most application layer research studies the resource allocation problem by considering either certain specific PHY/MAC schemes (e.g., IEEE 802.11 PHY) or some approximate relationship between resource and error probability. The advantages of considering an information-theoretic constraint at lower layers are two fold. First, the capacity region is an upper bound for the achievable rates where error-free transmission is possible. Since the capacity region is the fundamental characterization of the achievable rates, resource allocation with the capacity region constraint actually gives the *achievable video quality region*, which is the upper bound of achievable video quality. Second, using capacity region instead of any specific PHY/MAC scheme is more robust in the sense that the derived solution is insensitive to what specific scheme is actually used. Thus, the proposed policy can be adopted by a wide range of applications.

The existing literature on multi-user resource allocation has mainly focused on addressing the aforementioned two aspects in isolation. For instance, there are information-theoretic studies characterizing the *throughput* at the PHY/MAC. The entire throughput capacity region of a multiaccess fading channel is explicitly characterized in [1], [2]. Then the results are generalized to broadcast fading channels in [3]. It is generally believed that weighted-sum-rate-maximizing (WSRM) is the optimal operating point of the system. There are also studies addressing fairness [4], scheduling [5] and QoS [6]. In the video resource allocation problem, for example, a lot of research is devoted to studing the problem with specific lower layer schemes, e.g., IEEE 802.11 wireless LAN [7] or some given adaptive techniques [8].

In this paper, we will first show that the previous information-theoretic results which maximize the weighted sum rate of all users are suboptimal from the video performance perspective. We proceed to develop optimal resource allocation policies for several different models. We adopt operational video Quality-Rate (Q-R) models, in which the video quality is a strictly concave function of the transmission rate. For wireless communication of the video data, we consider three multiaccess capacity regions, each corresponding to a different type of channels:

a) *Non-fading AWGN channel*, and *fading channel with a given power control policy*. The capacity regions of both these channels exhibit the polymatroid structure, and we propose the resource allocation scheme that maximizes the weighted sum of video qualities of all

Manuscript received March 29, 2007; revised July 2, 2007, September 13, 2007, and November 11, 2007; accepted December 16, 2007. The associate editor coordinating the review of this paper and approving it for publication was D. Kim. This research was supported in part by NSF CAREER Award CCF 0541867 and UC Micro. The material in this paper was presented in part at the IEEE International Conference on Communications (ICC), Glasgow, Scotland, June 2007.

C. Shen and M. van der Schaar are with the Electrical Engineering Department, University of California Los Angeles (UCLA), Engineering IV Building, 420 Westwood Plaza, Los Angeles, CA 90095 (e-mail: {congshen, mihaela}@ee.ucla.edu).

Digital Object Identifier 10.1109/TWC.2008.070337.

users for any rates inside the corresponding multiaccess capacity region. We name the optimal policy *Largest Quality Improvement Highest Possible Rate (LQIHPR).* This policy has a low-complexity greedy algorithm: transmit an incrementally larger amount of video bits/packets until the capacity region constraint is tight. We further generalize LQIHPR to various video Q-R models. We prove the optimality of the generalized LQIHPR, and provide a "horizontal water-filling" greedy implementation algorithm. Numerical examples are shown to demonstrate the efficiency of the proposed algorithm.

b) *Fading channel with dynamic power control policy*. As shown in [2], the capacity is increased (due to the dynamic power allocation) compared to the transmission scenario in a). The optimal resource allocation solution is obtained by a "divide and conquer" strategy: divide the entire throughput capacity region into sub-regions and apply an modification to the Tse-Hanly solution [2]. We provide a low-complexity method to construct the candidate intervals in which the modified Tse-Hanly solution is applied. Since this "divide and conquer" solution is only a theoretical upper bound instead of an implementable algorithm, we do not provide numerical examples, but the optimality of the proposed solution is rigourously proved.

Our work differs from the previous information-theoretic studies, where the target is to determine the highest possible transmission rate that the physical channel can support given a power budget. Thus, it aims at determining the capacity region, without considering its usage for multimedia users. Our work, however, is devoted to answering the following question: For a given set of resources (power and rates), how should they be optimally allocated among multiple video users? The multiaccess capacity region only provides the constraint, i.e., it only tells us how many resources we have, but it does not determine how to optimally allocate the available resources among the users. Our proposed solution explicitly considers the operational rate-quality performance of the deployed video coders and, based on this, determines the optimal allocation according to a predetermined performance metric.

The rest of this paper is organized as follows. Section II defines the system model, briefly discusses previous results, and formulates the problem. Section III presents the LQIHPR policy together with a low-complexity greedy algorithm for non-fading channel and fading channel with any given power control policy. For dynamic power and rate allocation, the optimal policy is derived in Section IV. Finally, Section V concludes the paper.

II. System model and problem formulation

A. Channel Model

We adopt the same channel model as in [2]. Specifically, we consider an I-user Gaussian multiaccess fading channel with bandwidth W. The discrete-time channel model used in this paper is

$$Y(n) = \sum_{i=1}^{I} \sqrt{H_i(n)} X_i(n) + W(n) \tag{1}$$

where $X_i(n)$ and $H_i(n)$ are the transmitted symbol and the flat-fading process of user i at time n, respectively. $W(n)$ is the receiver additive white Gaussian noise (AWGN) with variance $N_0/2$ per dimension. Each user i is subjected to a long-term average power constraint: $E[|X_i(n)|^2] \leq \bar{P}_i$. A power allocation policy is called *feasible* if it satisfies this constraint. The time-varying fading processes $\{H_i(n), i = 1, \cdots, I\}$ are assumed to be jointly stationary and ergodic, and the channel coherent time is sufficiently large such that H_i can be considered constant over a long block length. We assume that the fading processes of different users are independent of each other.

B. Capacity Regions and the WSRM Policy

As in [2], we consider the case where both the receiver and the transmitters know CSI perfectly. Resource allocation is done by a central controller at the access point which takes the joint fading state $\mathbf{h}$ as an input, and outputs the power allocation $\mathcal{P}(\mathbf{h}) = (P_1(\mathbf{h}), \cdots, P_I(\mathbf{h}))$ and rate allocation $\mathcal{R}(\mathbf{h}) = (r_1(\mathbf{h}), \cdots, r_I(\mathbf{h}))$.

In this paper, the following three multiaccess capacity regions are considered.

1) For the Gaussian non-fading multiaccess channel $\mathbf{h}$ with constant transmit power $\mathbf{P}$, the capacity region is well known [9] to be

$$C_g(\mathbf{h}, \mathbf{P}) = \{\mathbf{r} : r(S) \leq C_S(\mathbf{h}, \mathbf{P}), \forall S \subseteq \{1, \cdots, I\}\} \tag{2}$$

where $C_S(\mathbf{h}, \mathbf{P}) \triangleq W \log\left(1 + \frac{\sum_{i \in S} h_i P_i}{N_0 W}\right)$, and $r(S)$ is used to denote $\sum_{i \in S} r(i)$ throughout this paper.

2) For the time-varying Gaussian multiaccess fading channel and a given feasible power allocation policy $\mathcal{P}(\mathbf{H})$, the capacity region has been proved [10] to be

$$\begin{aligned} C_f(\mathcal{P}) = \{\mathbf{r} : \mathbf{r}(S) &\leq \mathbb{E}_{\mathbf{H}}[C_S(\mathbf{H}, \mathcal{P}(\mathbf{H}))], \\ \forall S &\subseteq \{1, \cdots, I\}\}. \end{aligned} \tag{3}$$

Both (2) and (3) exhibit the polymatroid structure [2]. As will be discussed later, this leads to the same type of optimal rate allocation policies. For the sake of notation, (2) (3) are referred as *polymatroid-type regions*.

3) For the time-varying Gaussian multiaccess fading channel with dynamic power allocation, the capacity region is shown to be

$$C(\bar{\mathbf{P}}) = \bigcup_{\mathcal{P} \in \mathcal{F}} C_f(\mathcal{P}) \tag{4}$$

where $\mathcal{F}$ is the set of all feasible power allocation policies. This capacity region has been explicitly characterized in [2]. We refer to the capacity region (4) as a *convex-type region*.

From an information-theoretic viewpoint, it is generally of interest to maximize the weighted sum rate of users [1], [2]

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \boldsymbol{\mu}^T \mathbf{r} \\ \text{subject to} & \mathbf{r} \in \mathcal{C} \end{array} \tag{5}$$

where $\mathbf{r}$ is the rate vector, and $\boldsymbol{\mu}$ is the nonnegative weight vector with $\sum_{i=1}^{I} \mu_i = I$. Solutions to problem (5) have been derived for the capacity regions $C_g(\mathbf{h}, \mathbf{P})$, $C_f(\mathcal{P})$, and $C\left(\bar{\mathbf{P}}\right)$. Details can be found in [2] and the references therein. Such solutions are generally believed to be the optimal operating point of the system. We refer to the general solution to (5) as the *Weighted-Sum-Rate-Maximizing (WSRM)* policy.

C. Quality-Rate Models

Operational Q-R models [11], [12] describe the achievable quality of a specific video coder (e.g., H.264/AVC, MPEG-2, MPEG-4 or a 3D wavelet video coder) as a function of the allocated rate. We use the widely accepted Peak Signal-to-Noise Ratio (PSNR) as a measure of video quality. The operational Q-R model adopted in this work is also a widely used one [11] where for user i we use N_i line segments with slopes $\lambda_i^{(k)}, k = 1, 2, \cdots, N_i$, each of which corresponds to a rate interval $\Delta_i^{(k)}$:

$$Q_i(r_i) = \begin{cases} 0, & r_i < r_i^{min} \\ q_i^{(k)} + \lambda_i^{(k)}\left(r_i - r_i^{(k)}\right), & r_i \in \Delta_i^{(k)}, 1 \leq k \leq N_i \end{cases} \tag{6}$$

where $(r_i^{(k)}, q_i^{(k)}), k > 1$ is the connection point of two line segments, and $(r_i^{(1)}, q_i^{(1)}) = (r_i^{min}, q_i^{min})$ represents the minimum rate-quality requirement. This is based on the observation that below this point, the transmission results in unacceptable video quality and hence, a user that cannot obtain his minimum quality will not participate in the wireless transmission. It is a well known result (see e.g. [11]) that $\lambda_k \geq \lambda_{k+1}$, i.e., the quality benefit derived by an operational video coder decreases with an increased allocated source rate.

In practical video coders, the Q-R model is generally discrete. Our work mainly focuses on the Q-R model with fine granularity. In this case, the continuity of Q-R model can be considered as an ideal case, because how continuous/discrete the Q-R model is depends on the granularity of the video packets. As the granularity of the video packets becomes finer, the overall Q-R function becomes less discrete. Also, the practical discreteness only degrades the performance compared to the continuous case. However, for the highly discrete models, the continuous relaxation might lead to severe performance degradation. In this case the proposed method is not applicable, and one has to resort to numerical discrete optimizations.

D. The Weighted-Sum-Quality-Maximizing Policy

Multimedia transmission applications aim at directly optimizing the APP layer utility (i.e., video quality). Thus, the focus of this paper is on how to allocate the resources (power and rate) to different users such that the weighted sum of video qualities is maximized. This can be formally cast as

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \sum_{i=1}^{I} w_i Q_i(r_i) \\ \text{subject to} & \mathbf{r} \in \mathcal{C} \end{array} \tag{7}$$

where $w_i \geq 0$ and $\sum_{i=1}^{I} w_i = I$. The weights $w_i, i = 1, \cdots, I$ are used as an adjustment from system consideration, e.g., to deal with the asymmetric channel conditions of different users, or to implement some fairness control. We will refer to the solution to this general optimization problem as the *Weighted-Sum-Quality-Maximizing (WSQM)* policy. In this paper, we focus on the wireless multiaccess fading channel, and derive the WSQM policies for

1) $Q(r)$ using line-segment model (6), or any other R-D models with packets prioritization mechanism;
2) $\mathcal{C}$ equal to $C_g(\mathbf{h}, \mathbf{P})$, $C_f(\mathcal{P})$, and $C\left(\bar{\mathbf{P}}\right)$, respectively.

Note that the aforementioned WSRM solution is generally suboptimal for video applications. This is because even if two users are assigned the same rate, their video quality might differ significantly, due to the nonlinear relationship between video quality and rate. This is the motivation for WSQM.

III. Optimal resource allocation for polymatroid-type capacity regions

We first study problem (7) where $\mathcal{C} = C_g(\mathbf{h}, \mathbf{P})$ or $\mathcal{C} = C_f(\mathcal{P})$. Both capacity regions have a polymatroid structure, and thus they will have similar solutions, as will be evident in this section. For the sake of notation simplification, we only use $C_g(\mathbf{h}, \mathbf{P})$ in the following derivation, i.e., we will solve

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \sum_{i=1}^{I} w_i Q_i(r_i) \\ \text{subject to} & \mathbf{r} \in C_g(\mathbf{h}, \mathbf{P}) \end{array} \tag{8}$$

in this section. It should be noted the same method can be applied to $C_f(\mathcal{P})$ without any modification.

A. Largest Quality Improvement Highest Possible Rate

First, we need to determine in which area of the multiaccess capacity region the optimal solution is. For this, we cite the definition of *boundary surface* from [2, Definition 3.9].

Definition 1: The boundary surface of the multiaccess capacity region $C_g(\mathbf{h}, \mathbf{P})$ (or $C_f(\mathcal{P})$, $C\left(\bar{\mathbf{P}}\right)$) is the set of rates such that no component can be increased with the other components remaining fixed while in the capacity region.

For example, in a two-user scenario, the boundary surface of $C_g(\mathbf{h}, \mathbf{P})$ is given by the line segment $\{(r_1, r_2) : r_1 + r_2 = C_{\{1,2\}}(\mathbf{h}, \mathbf{P}), r_i \leq C_{\{i\}}(h_i, P_i), i = 1, 2\}$.

The following theorem gives the possible locations of the optimal operating point within the general multiaccess capacity region $\mathcal{C}$.

Theorem 1: The solution to (7) must be at the boundary surface of the multiaccess capacity region $\mathcal{C}$.

Proof: Due to the fact that the Q-R function is monotonically nondecreasing and according to the definition of boundary surface, the proof directly comes from the *Pareto Optimality* [13]. ■

The remaining problem is how to find the operating point at the boundary surface. We propose a low-complexity greedy algorithm to solve this problem. First, we can incorporate the weight w_i into the slopes of each user's Q-R function. Thus, without loss of generality we assume all w_i to be equal to 1 for the remaining of this paper. Second, we assume throughout this paper that each user has an individual rate limit which is larger than the minimum rate required in its Q-R model: $C_{\{i\}}(\mathbf{h}, \mathbf{P}) > r_i^{min}$. This is reasonable because otherwise we can always allocate zero rate to the user and exclude him from the following algorithms. We also assume that the users who participate in the resource allocation have their rates and video PSNRs in similar ranges. This can be done by setting some threshold and excluding disqualified users. The reason for this assumption is to avoid extreme situations such as one user's rate range is much higher than the others, or his video PSNRs are much lower.

Algorithm 1 is the proposed solution to problem (8). We name it *Largest Quality Improvement Highest Possible Rate (LQIHPR)*, as we always increase the rate of the user who has the steepest quality improvement. The proof of the optimality of Algorithm 1 is deferred to Section III-B, where we prove a more general result. The complexity is roughly linear in the number of active users, depending on the number of rate intervals in the Q-R models.

Algorithm 1 I-User Greedy Rate Allocation Algorithm for Line-segment Q-R Models

```
Input: $C_g(\mathbf{h},\mathbf{P})$ (2); User $i$'s Q-R model (6) with
  slope set $\left\{\lambda_i^{(1)},\lambda_i^{(2)},\cdots,\lambda_i^{(N_i)}\right\}$ and rate interval set
  $\left\{\Delta_i^{(1)},\Delta_i^{(2)},\cdots,\Delta_i^{(N_i)}\right\}$, $i=1,\cdots,I$.
Initialization: Sort the slopes from all users
  $\left\{\lambda_i^{(1)},\lambda_i^{(2)},\cdots,\lambda_i^{(N_i)}\right\}_{i=1}^{I}$ in descent order and form the
  ordered slope set $\mathbf{\Lambda}_{order}=\left\{\cdots\geq\lambda_{j_1}^{(k_{j_1})}\geq\lambda_{j_2}^{(k_{j_2})}\geq\cdots\right\}$
  with the corresponding rate interval set $\mathbf{\Delta}_{order}=$
  $\left\{\cdots,\Delta_{j_1}^{(k_{j_1})},\Delta_{j_2}^{(k_{j_2})},\cdots\right\}$; Allocate user $i$ with an initial
  rate $r_i=r_i^{min}$, $i=1,\cdots,I$.
Repeat:
  1) Select the first available slope $\lambda_j^{(k)}$ from the ordered
     slope set $\mathbf{\Lambda}_{order}$, and determine the corresponding user
     $j$;
  2) Increase the rate $r_j$ of user $j$ until the rate interval $\Delta_j^{(k)}$
     is fulfilled, or any rate limit is reached;
  3) Delete $\lambda_j^{(k)}$ / $\Delta_j^{(k)}$ from sets $\mathbf{\Lambda}_{order}$ / $\mathbf{\Delta}_{order}$. In case
     that any rate limit is reached, delete all remaining
     slopes/rate intervals associated with the corresponding
     user(s) from sets $\mathbf{\Lambda}_{order}$ / $\mathbf{\Delta}_{order}$.
Until: $\mathbf{\Lambda}_{order}$ / $\mathbf{\Delta}_{order}$ is empty, or the overall $I$-user sum
  rate limit is reached.
Return: $\mathbf{r}^*=(r_1,\cdots,r_I)$.
```

Now we try to illustrate this algorithm using a two-user case. From Theorem 1, the solution to

$$\begin{array}{ll}\underset{r_1,r_2}{\text{maximize}} & Q_1(r_1)+Q_2(r_2)\\ \text{subject to} & \mathbf{r}\in C_g(\mathbf{h},\mathbf{P})\end{array}$$

must lie in the line segment $\{(r_1,r_2) : r_1+r_2 = C_{\{1,2\}}(\mathbf{h},\mathbf{P}), r_i\leq C_{\{i\}}(h_i,P_i), i=1,2\}$. By noticing that the slopes of each user's Q-R model are monotonically decreasing as its rate increases, a typical ordered slope set $\mathbf{\Lambda}_{order}$ could be $\mathbf{\Lambda}_{order} = \left\{\lambda_1^{(1)}\geq\lambda_2^{(1)}\geq\lambda_2^{(2)}\geq\lambda_1^{(2)}\geq\cdots\right\}$, and the associated rate interval set is $\mathbf{\Delta}_{order} = \left\{\Delta_1^{(1)},\Delta_2^{(1)},\Delta_2^{(2)},\Delta_1^{(2)},\cdots\right\}$.

Fig. 1 gives two examples showing how the greedy rate allocation is performed based on the scenario described in the previous paragraph. Rate is allocated to users according to their slopes' ordering. In Fig. 1 (a), user 1 and 2 increase their rate in the order of $1,2,2,1,2$ until user 1 first stops at its individual maximum rate $C_{MAX}(\{1\})$, and then user 2 continues being allocated more rate until the maximum sum rate limit $C_{MAX}(\{1,2\})$ is reached. The example in Fig. 1 (b) shows another possibility that neither user's individual rate limit is reached, but the sum rate limit $C_{MAX}(\{1,2\})$ is met. In this situation the optimal operating point is not at any extreme point but strictly inside the line segment. Again this demonstrates that the conventional WSRM policy which operates at one extreme point is suboptimal for video allocation.

Since LQIHPR and WSRM generally operate at different points in the boundary surface, the methods to achieve them are also different. As we have mentioned before, to maximize the weighted sum rate one has to operate at a specific extreme point of the capacity region, and *successive decoding* is sufficient to achieve the corresponding rate pair. In LQIHPR we generally operate within the boundary surface, which means *time sharing* is necessary[1].

B. LQIHPR for General Q-R Models

At this moment it seems that the optimality of the proposed LQIHPR policy relies on the line-segment Q-R model (6). However, we argue that the optimality is not depending on any specific Q-R model being used. Instead, LQIHPR can be proved to be optimal for *any* utility-rate model that is monotonically increasing and strictly concave. In this subsection, we will prove that LQIHPR is optimal for such a broad class of Q-R models.

To be more specific, let us look at problem (8) from a different perspective. We do not consider any specific Q-R model $Q(r)$. Instead, we make two reasonable assumptions about $Q(r)$:

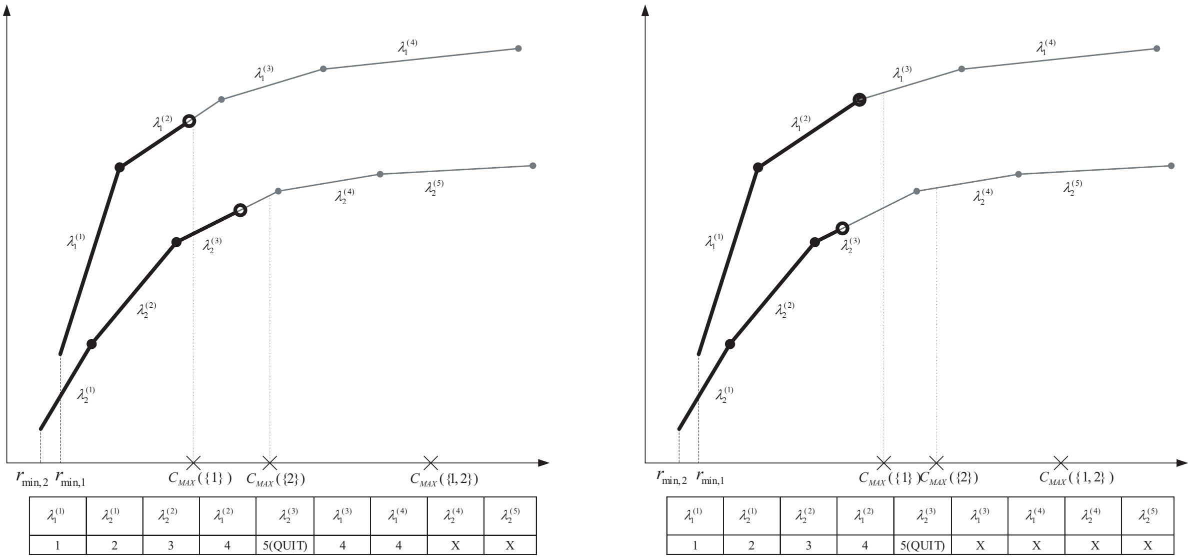
1) function $Q(r)$ is continuously differentiable, or has a finite set of nondifferentiable points, and thus it has a first-order derivative function $f(r)\triangleq \mathrm{d}Q(r)/\mathrm{d}r$ which has at most finite discontinuous points.
2) the derivative function $f(r)$ is nonincreasing.

The first assumption holds for state-of-the-art video coders designed for wireless streaming applications (e.g., H.264, MPEG-4 Fine Granularity Scalability (FGS) and wavelet based video coders), as they need to provide the ability to efficiently adapt to dynamic changes in the channel conditions [15, Chapter 4]. Hence, they have the ability to gracefully increase the received video quality for every small increase in source rate by successively refining the source information. This is implemented in MPEG-4 FGS and wavelet video coders using embedded quantization and in H.264 using flexible data partitioning. The second assumption has been justified in [11].

With these two assumptions, $Q(r), r\geq r^{min}$ can be written as

$$Q(r)=\int_{r^{min}}^{r} f(z)\mathrm{d}z$$

[1] In general there are other ways to achieve points strictly inside the boundary surface other than time-sharing, e.g., the rate-splitting approach [14]. Here the terminology "time sharing" is used in a more general sense.



$\lambda_1^{(1)}$	$\lambda_2^{(1)}$	$\lambda_2^{(2)}$	$\lambda_1^{(2)}$	$\lambda_2^{(3)}$	$\lambda_1^{(3)}$	$\lambda_1^{(4)}$	$\lambda_2^{(4)}$	$\lambda_2^{(5)}$
1	2	3	4	5(QUIT)	4	4	X	X

(a) Individual rate limit is reached.

$\lambda_1^{(1)}$	$\lambda_2^{(1)}$	$\lambda_2^{(2)}$	$\lambda_1^{(2)}$	$\lambda_2^{(3)}$	$\lambda_1^{(3)}$	$\lambda_1^{(4)}$	$\lambda_2^{(4)}$	$\lambda_2^{(5)}$
1	2	3	4	5(QUIT)	X	X	X	X

(b) Sum rate limit is reached.

Fig. 1. Two examples illustrating Algorithm 1. In the table below each plot, the first row shows the slope ordering, and the second row indicates the deletion ordering as in repeat step 3) of Algorithm 1. Here $C_{MAX}(S)$ means the sum rate constraint for the entire set S is tight.

and problem (8) is equivalent to

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \sum_{i=1}^{I} \int_{r_i^{min}}^{r_i} f_i(z)\mathrm{d}z \\ \text{subject to} & \mathbf{r} \in C_g(\mathbf{h}, \mathbf{P}). \end{array}$$

Thus, problem (8) can be reinterpreted as maximizing the sum area under each $f_i(z)$ subject to the rate constraints. With this interpretation, we can explain why LQIHPR is always optimal as long as $Q(r)$ used in problem (8) satisfies the two assumptions, using a *quality-splitting* approach which is similar to the *rate-splitting* explanation in [2]. Let us define $f_i(z)$ as the marginal quality function for user i, and $f_i(z)\mathrm{d}z$ can be interpreted as the marginal increase in video quality of user i due to allocating rate $\mathrm{d}z$ to user i at the rate level z. Then at the rate level $\mathbf{z}$, the optimal solution can be obtained by always allocating rate $\mathrm{d}z$ to the users where it leads to the maximum marginal increase in video quality. Fig. 2 shows a three-user example. Details of this example will be explained later. In summary, we give the LQIHPR algorithm for general Q-R models in Algorithm 2.

Theorem 2: Algorithm 2 gives the optimal solution to problem (8).

Proof: See Appendix A. ■

Algorithm 2, similar to Algorithm 1, is a LQIHPR solution: rate is always given to the user(s) for which it leads to the maximum quality increase. We further point out that this is conceptually a "*horizontal water-filling*" process. We will explain this idea using the three-user example in Fig. 2. Initially each user is assigned with the minimal rate r_i^{min}. A horizontal water-filling starts from point A since that is the highest position among all $f_i(r_i)$'s. Water horizontally (from left to right) fills in the vessel under $f_2(r)$ until it hits the level of the second highest point B. Then besides filling the vessel under $f_2(r)$, water also pours into the vessel under $f_1(r)$ and thus it will simultaneously fill in both vessels. Similarly it continues until water level C is reached, where user 3 is added into this horizontal water-filling process. Then when water comes to level D where user 2's individual limit is reached, water-filling for user 2 stops. Remaining user 1 and 3 continue filling water until the maximum limit for all users is met, and they stop at water level E/F.

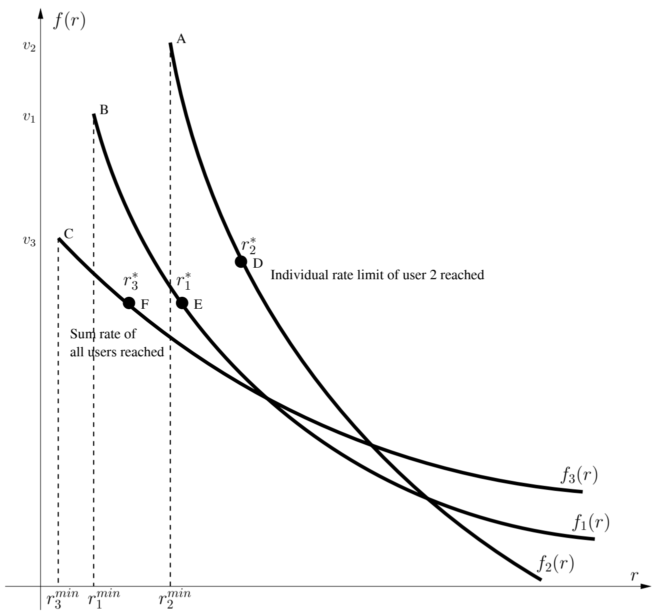


Fig. 2. A three-user example illustrating the LQIHPR rate allocation algorithm for general Q-R models.

It is important to emphasize that there are both connections and important differences between the "horizontal water-filling" algorithm here and the conventional multi-user water-filling solution in [1], [2]. The connection is that both solutions have a similar flavor: always allocate rate to those who have the steepest ascent of the objective function. The differences are multifold. First of all, in our problem setting we are actually doing water-filling in the *utility domain*, where the objective functions are *concave* in rates, while

Algorithm 2 I-User Greedy Rate Allocation Algorithm for General Q-R Models

```
Input: $C_g(\mathbf{h}, \mathbf{P})$ (2); User $i$'s Q-R model $Q_i(r_i)$ which must
    be a continuous (or has a finite set of discontinuous points)
    function with nonincreasing derivatives, $i = 1, \cdots, I$.
Initialization:
    1) Allocate user $i$ with an initial rate $r_i = r_i^{min}$.
    2) Calculate the derivative function $f_i(r) = \frac{dQ(r)}{dr}$.
    3) Calculate $v_i = f_i(r_i^{min}), i = 1, \cdots, I$ and form an
       ordered set of $v_i$: $\{v_{i_1} \geq \cdots \geq v_{i_I}\}$. Form active user
       set $U_a = \{i_1\}$, active user rate set $R_a = \{r_{i_1}\}$, and
       active marginal quality set $F_a = \{f_{i_1}(r_{i_1})\}$.
    4) $k = 1$.
Repeat:
  Simultaneously increase all the rates in $R_a$ until
  a) a subset $U_s \subset U_a$ saturates their sum rate limit $r(U_s) =$
  $C_{MAX}(U_s)$. Then denote $R_s, F_s$ as the active rate and
  marginal quality set corresponding to $U_s$, respectively, and
      if $U_s = \{1, \cdots, I\}$ then
          exit loop.
      else
          remove $U_s$ from $U_a$, $R_s$ from $R_a$, $F_s$ from $F_a$, and
  continue increasing the rates of users in $U_a$.
      end if
  b) $f_{i_1}(r_{i_1}) = \cdots = f_{i_k}(r_{i_k}) = v_{i_{k+1}}$. Then
          add $i_{k+1}$ into $U_a$, $r_{i_{k+1}}$ into $R_a$, and $f_{i_{k+1}}(r_{i_{k+1}})$ into
  $F_a$. $k = k + 1$.
Return: $\mathbf{r}^* = (r_1, \cdots, r_I)$.
```

the conventional information-theoretic water-filling is in the *rate domain* with *linear* objective functions. Our solution is more general than the conventional one, since one can always view linearity as a special case of concavity. Secondly, the "greedy" behavior in our water-filling solution is different than in the conventional one. In multi-user water-filling, the greedy algorithm is strongly *competitive*: it picks up the largest marginal utility function at each interference level $N_0 + z$. In other words, each time there is at most one user being chosen. However, in our solution, the greedy behavior is *collaborative*: we look at the marginal utility functions of *all* users all the time, and multiple users can be chosen simultaneously, as long as they have the same steepest ascent.

As we have stated at the beginning of this subsection, the specific Q-R model is not fundamental in deriving the optimal solution. To better understand this, it would be interesting to see how Algorithm 1 can be regarded as a special case of Algorithm 2. Fig. 3 illustrates the corresponding line-segment Q-R model for the same three-user example as in Fig. 2. According to Algorithm 2, since level A is the highest one, we will first increase the rate of user 2 until level A drops to level E. Then the second highest level B will make the rate of user 1 increase, and then the third highest level C makes the rate of user 1 further increase, and then level D for user 3, and so on. This is actually equivalent to Algorithm 1. However there are some differences between these two algorithms. For a line-segment model, the water level keeps constant inside one rate interval, and changes dramatically at the edge, which is a nondifferentiable point in $Q(r)$. This makes the water filling generally only pour into one user's vessel at a time, and change to another user at the edge. The general Q-R model, on the other hand, has a gradually changing shape, and thus water can pour into several users simultaneously, in a gradual fashion.

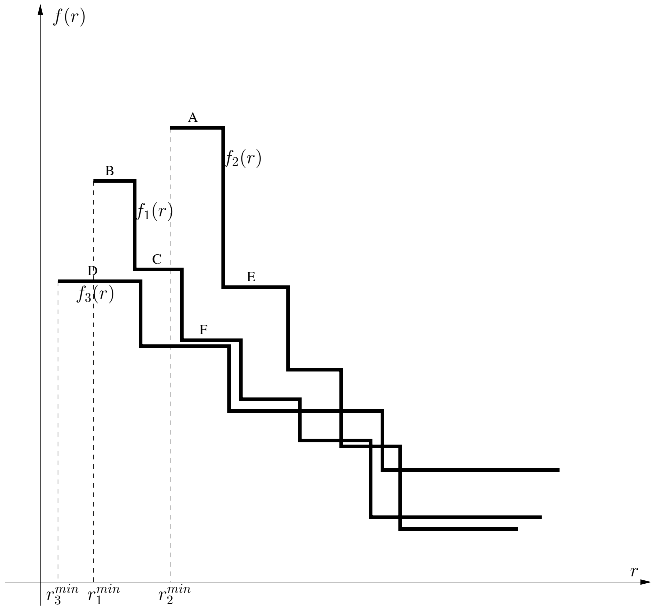


Fig. 3. Illustration of how Algorithm 1 is a special case of Algorithm 2 with the same three-user example as in Fig. 2.

The key reason that makes the LQIHPR policy optimal is the *monotonically decreasing slope* property of the Q-R model. Due to prioritizing bits/packets according to their descent impact on the overall video quality, the increase of quality will become smaller as the rate increases [11]. This directly translates to an ever-decreasing slope in the Q-R function. It is important to notice that all the video coders for wireless video applications designed so far generate a Q-R function with decreasing slopes by *bits/packets prioritization*, and thus the optimality of LQIHPR is universal to all video coding schemes. For more information on various prioritization schemes for hybrid video coders and wavelet coders, the reader is referred to [16] and [17], respectively.

C. Numerical Examples

We have theoretically demonstrated the optimality of the proposed LQIHPR (WSQM) in terms of maximizing video quality-rate functions. In this section, we provide several sets of simulation results to access the performance difference between LQIHPR and WSRM policies. It is notable that the WSRM policy does not have access to the quality-rate information, and thus is certainly sub-optimal in terms of video quality. However, the numerical comparison made in this section helps quantify how much gain we can get by performing the resource allocation based on knowing the quality-rate model, and especially when significant improvement in video quality may be achieved.

We provide five sets of simulations, considering different channel conditions, different number of users/videos, and different delay constraints on video coders. In the first four sets of simulations, we assume that user 1 wants to transmit the *Mobile* video, user 2 has the *Coastguard* video for transmission, and user 3 transmits *Stefan*. The videos used throughout all simulations are standard ones in the video coding community. H.264/AVC encoder is used throughout all simulations to compress the videos with the following parameters: one B-frame, GOP size 16 video frames, maximum 2 reference frames for motion compensation/estimation (previous I-frames or P-frames), worst end-to-end delay 66 ms, and CIF resolution 30 Hz. Line-segment Q-R model (6) is used throughout all simulations, where the average PSNR reported for the Y channel from the H.264/AVC encoder is used. In the line-segment Q-R model, there is not only a minimum rate constraint, but also a maximum rate limit. These limits are summarized in Table I. Thus if the rate assigned to one user is larger than the upper limit, the quality will stop increasing. LQIHPR is implemented with all weights equal to 1 and compared with WSRM. As we have proved, each extreme point of the multiaccess capacity region is an optimal solution of WSRM for certain weights. We simply choose one out of all the $I!$ extreme points which gives the largest sum of video qualities. Notice that this is the highest sum video quality WSRM can provide.

We first consider a two-user Rayleigh fading symmetric multiaccess channel with average channel power (Table II). Each user is assumed to have an average receive SNR of 10 dB and bandwidth 1 MHz. This bandwidth is used throughout this section. Simulation shows that the LQIHPR policy, which aims at maximizing the sum quality, has an average sum PSNR of 71.8 dB, while the WSRM policy only provides 69.3 dB: there is an approximately 2.5 dB average quality gain in this simulation setting. It is interesting to explore how the WSRM policy results in such a suboptimal video performance. WSRM always operates at one extreme point, where only one user gets its maximum rate, and thus its best possible video quality. However, this "selfish" allocation leaves very small room for the other user to increase its quality. Heuristically, if the first user can give some rate to the second one without decreasing the sum rate, he might experience very limited quality drop, but at the same time the video quality of the second user might increase significantly due to the nonlinear relationship between rate and quality, and thus the sum of video qualities can be increased.

The second simulation has the same environment as the first one, except that the channels are asymmetric (Table II). We assume that user 1 has an average SNR of 10 dB, while user 2 has 15 dB. This models the situation where one user has a better channel than the other. In this simulation, WSRM gives an average PSNR of 66.5 dB, and LQIHPR provides 73.9 dB. There is a 7.4 dB performance difference. It can be noted that in this asymmetric situation the benefit of LQIHPR is even larger, which can be explained as following. The fact that user 2 has higher receive SNR means user 2 will typically have a better channel than user 1. Translating into the multiaccess capacity region, the rate of user 2 is much larger than that of user 1 at the boundary surface. In this region, typically the video quality of user 2 has already saturated, while user 1 may operate around its minimum rate-quality point such that its video quality can be significantly improved by a very small rate increase. In other words, the same rate is much more important to user 1 than to user 2 in terms of video quality. WSRM operates at the extreme point where the rate of user 2 is maximized, while LQIHPR, by noticing the fact that rate is more important to user 1 than to user 2, decreases the rate of user 2 and allocates it to user 1, and maximizes the sum of video qualities.

From these simulations results we can conclude under what conditions the resource allocation gain will be large. Typically this happens when one user is operating at the "saturation regime" while another user is at the "survival regime". In such case significant performance gain can be achieved if the first user gives some of his rate to the second user such that without degrading his performance much, the second user can benefit significantly. On the other hand, the resource allocation gain can be very limited if both users are at the "saturation regime", where video qualities are insensitive to small rate changes.

The third and fourth simulations include three users (Table III). In the third simulation we simulate a Rayleigh fading symmetric multiaccess channel, with the same parameters as in the first simulation. LQIHPR results in an average sum PSNR of 105.8 dB, while WSRM gives 104.1 dB. The last simulation includes a three user asymmetric Rayleigh fading multiaccess channel. User 1 has an average receive SNR 15 dB, while user 2 and 3 have 18 and 13 dB, respectively. The other parameters are the same as before. In this setting LQIHPR gives an average sum PSNR 111.5 dB, compared with 107.1 dB from WSRM. The conclusions from these three-user simulations comply with those two-user examples.

The last simulation is developed to deal with delay constraint and its impact on the resource allocation policy. In this simulation, both user 1 and user 2 want to transmit the *Akiyo* video to a central receiver. However, the H.264/AVC encoder generates different GOP structures to meet with the different worst-case end-to-end delay constraints of each user. User 1's worst-case end-to-end delay is 99 ms, while user 2's is 165 ms. The bandwidth is set to be 160 KHz, as the Akiyo video has very slow movement, and thus the required transmission rate is smaller than those three previous videos. The other parameters are the same as the previous four simulations. The results are reported in Table IV. It is clear that by properly modeling different delay constraints, the proposed LQIHPR is still efficient in terms of maximizing the video quality. As we have emphasized in Section III-B, as long as the quality-rate model satisfies the two mild assumptions, the LQIHPR policy is proved to be optimal.

IV. Optimal resource allocation for the convex-type capacity region

A. Problem Formulation

We have studied the problem that for any feasible power control policy $\mathcal{P}(\mathbf{H})$ that is given *a priori*, how to allocate the rate inside the resulting multiaccess capacity region to different users such that the weighted sum of video qualities is maximized. This problem setting, although has some implementation advantages such as low complexity due to a non-dynamic power allocation policy, does not make full use

TABLE I
MINIMUM AND MAXIMUM RATE/PSNR LIMITS. PSNR IN dB AND RATE IN MBPS.

	minimum rate	minimum PSNR	maximum rate	maximum PSNR
Mobile	0.60	28.3	3.30	38.1
Coastguard	0.22	28.1	2.50	37.8
Stefan	0.39	29.8	2.34	39.3

TABLE II
SIMULATION RESULTS FOR TWO-USER FADING MULTIACCESS CHANNELS. PSNR IN dB AND RATE IN MBPS.

	Policy	sum PSNR	PSNR 1	PSNR 2	sum rate	rate 1	rate 2
Symmetric MAC	LQIHPR	71.8	34.5	37.3	4.06	1.92	2.14
	WSRM	69.3	31.5	37.8	4.06	1.14	2.92
Asymmetric MAC	LQIHPR	73.9	36.1	37.8	5.00	2.50	2.50
	WSRM	66.5	28.7	37.8	5.00	0.66	4.34

TABLE III
SIMULATION RESULTS FOR THREE-USER FADING MULTIACCESS CHANNELS. PSNR IN dB AND RATE IN MBPS.

	Policy	sum PSNR	PSNR 1	PSNR 2	PSNR 3	sum rate	rate 1	rate 2	rate 3
Symmetric MAC	LQIHPR	105.8	34.5	34.6	36.7	4.72	1.92	1.26	1.54
	WSRM	104.1	37.1	31.8	35.2	4.72	2.90	0.66	1.16
Asymmetric MAC	LQIHPR	111.5	34.5	37.7	39.3	6.57	1.92	2.31	2.34
	WSRM	107.1	33.4	34.4	39.3	6.57	1.60	1.23	3.74

TABLE IV
SIMULATION RESULTS FOR TWO-USER FADING MULTIACCESS CHANNELS WITH DELAY CONSTRAINT. PSNR IN dB AND RATE IN MBPS.

Policy	sum PSNR	PSNR 1	PSNR 2	sum rate	rate 1	rate 2
LQIHPR	80.3	40.5	39.8	0.39	0.24	0.15
WSRM	75.1	41.4	33.7	0.39	0.33	0.05

of the flexibility in assigning power and rate, resulting in a capacity loss [2].

Now let us consider problem (7) with $\mathcal{C} = C\left(\bar{\mathbf{P}}\right)$ and $w_i = 1$:[2]

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \sum_{i=1}^{I} Q_i\left(r_i\right) \\ \text{subject to} & \mathbf{r} \in \mathcal{C}(\bar{\mathbf{P}}) \end{array} \tag{9}$$

where the Q-R model is given in (6).

At the first sight this problem is similar to (8): finding an operating point inside the capacity region to maximize the weighted sum of video qualities. However, this problem is essentially much harder due to the following reasons.

1) Problem (8) isolates the rate allocation problem from power allocation. It is assumed that a feasible power allocation scheme is chosen before studying the rate allocation problem. Problem (9), however, is a joint power and rate allocation one. Since the increase in capacity solely comes from the ability to allocate power dynamically [2], one cannot fix the power allocation any more: every point in the boundary surface of $C\left(\bar{\mathbf{P}}\right)$ is associated with a different feasible power control policy.
2) The explicit characterization of $C\left(\bar{\mathbf{P}}\right)$ is much more complicated than $C_g\left(\mathbf{h}, \mathbf{P}\right)$ or $C_f\left(\mathcal{P}\right)$. Since we are mainly interested in the boundary surface, let us make a comparison between the computation of boundary surfaces of $C_g\left(\mathbf{h}, \mathbf{P}\right)$ and $C\left(\bar{\mathbf{P}}\right)$. The boundary surface of $C_g\left(\mathbf{h}, \mathbf{P}\right)$ is a polyhedron, and thus several linear equations are sufficient to represent it. The boundary surface of $C\left(\bar{\mathbf{P}}\right)$, however, is a convex set, and for every point on the boundary surface there are two integral equations associated with it which have to be solved [2].

B. Boundary Surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$

We will solve problem (9) by first modifying the solution in [2], which was originally developed to characterize the boundary surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$, and then proposing a "divide and conquer" strategy. First of all, we qualitatively show where the solution might be, and justify the optimality of successive decoding.

Lemma 1: The solution to the optimization problem (9) must lie at the boundary surface of the capacity region $\mathcal{C}\left(\bar{\mathbf{P}}\right)$ in which the rate and power have a one-to-one mapping

$$\begin{aligned} r_{\pi(1)} &= \frac{1}{2}\log\left(1 + \frac{h_{\pi(1)}p_{\pi(1)}}{N_0}\right) \\ r_{\pi(k)} &= \frac{1}{2}\log\left(1 + \frac{h_{\pi(k)}p_{\pi(k)}}{N_0 + \sum_{i=1}^{k-1} h_{\pi(i)}p_{\pi(i)}}\right), k = 2, \cdots, I \end{aligned} \tag{10}$$

where π is a permutation on $\{1, \cdots, I\}$, which is determined by $\{Q_i(r_i), i = 1, \cdots, I\}$.

Proof: The first part is a direct result from Theorem 1. According to [2, Lemma 3.10], the boundary surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$ is the set of extreme points of $C_f\left(\mathcal{P}\right), \mathcal{P} \in \mathcal{F}$. Since the solution to problem (9) is at the boundary surface, it must be a successive decoding solution. ■

Remark 1: This lemma helps eliminate one degree of freedom in the optimization problem by revealing that the optimal solution $(\mathbf{r}^*, \mathbf{P}^*)$ must satisfy certain one-to-one mapping. However, the mapping itself depends on the Q-R model.

Remark 2: This lemma reveals an important difference from the LQIHPR policy. Recall that LQIHPR requires the system to operate in a *time-sharing* fashion. Equation (10), however, says that in the dynamic resource allocation problem the system is always working in a pure *successive decoding*

[2] As has been stated before, if $w_i \neq 1$ we can always incorporate them into the Q-R model.

fashion, i.e., time-sharing is not needed here. This is due to the fact that the boundary surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$ consists of pure successive decoding points.

From Lemma 1 we only need to consider the boundary surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$. Tse and Hanly provided a solution [2, Theorem 3.16] to explicitly characterize the boundary surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$. This solution is referred as the *Tse-Hanly solution* in the sequel. It is interesting to note that although the purpose of the Tse-Hanly solution is to explicitly characterize the entire multiaccess capacity region $\mathcal{C}\left(\bar{\mathbf{P}}\right)$, it turns out that this coincides with solving the optimization problem

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \boldsymbol{\mu}\mathbf{r} \\ \text{subject to} & \mathbf{r} \in C\left(\bar{\mathbf{P}}\right) \end{array} \tag{11}$$

for *all* possible $\boldsymbol{\mu} \in \mathbb{R}_+^I$. The key observation is that the objective function in (11) is linear in rate $\mathbf{r}$. We will develop a modification to the Tse-Hanly solution and apply it to solve our problem in the following.

C. A "Divide and Conquer" Solution

Now let us turn to problem (9). Since the Q-R model (6) for each user is a line-segment one, by an appropriate partitioning of the capacity region $\mathcal{C}\left(\bar{\mathbf{P}}\right)$ based on the length of the rate interval for each line segment, we can divide the original problem (9) into several parallel sub-problems, in each of which the objective function reduces to a linear combination of rate $\mathbf{r}$, and thus Tse-Hanly solution can be adopted to find the optimal solution for this sub-problem. To be more specific, since user $i, i = 1, \cdots, I$ is associated with $\left\{(\lambda_i^{(1)}, \Delta_i^{(1)}), \cdots, (\lambda_i^{(N_i)}, \Delta_i^{(N_i)})\right\}$, we can partition the entire rate of interest into I-dimensional rate hypercubes $\boldsymbol{\Delta}^I \triangleq \left\{\left(\Delta_1^{(k_1)}, \Delta_2^{(k_2)}, \cdots, \Delta_I^{(k_I)}\right), \forall k_i = 1, \cdots, N_i, i = 1, \cdots, I\right\}$, and apply this partition to the capacity region $\mathcal{C}\left(\bar{P}\right)$. There is also a set of slope vectors $\boldsymbol{\Lambda}^I \triangleq \left\{\left(\lambda_1^{(k_1)}, \lambda_2^{(k_2)}, \cdots, \lambda_I^{(k_I)}\right), \forall k_i = 1, \cdots, N_i, i = 1, \cdots, I\right\}$, each element in which corresponds to one hypercube in $\boldsymbol{\Delta}^I$. A geometric illustration of this idea for a two-user situation is given in Fig. 4.

Let us temporarily assume that we know the subset $\boldsymbol{\Delta}_{bs}^I \subset \boldsymbol{\Delta}^I$ which is the set of all the I-dimensional hypercubes that fully covers the boundary surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$. We name them "active hypercubes" since they are the ones that will be used in applying Tse-Hanly solution. In the two-user example as in Fig. 4, this is the set of all the gray rectangles. The cardinality of $\boldsymbol{\Delta}_{bs}^I$ is assumed to be M, and we represent this set as

$$\begin{aligned} \boldsymbol{\Delta}_{bs}^I &= \left\{ \left(\Delta_{bs}^{(k_{1,1})}, \cdots, \Delta_{bs}^{(k_{I,1})}\right), \cdots, \right. \\ &\qquad \left. \left(\Delta_{bs}^{(k_{1,M})}, \cdots, \Delta_{bs}^{(k_{I,M})}\right) \right\} \\ &\triangleq \left\{\boldsymbol{\Delta}_{bs}^1, \cdots, \boldsymbol{\Delta}_{bs}^M\right\} \end{aligned}$$

with the corresponding slope set

$$\begin{aligned} \boldsymbol{\Lambda}_{bs}^I &= \left\{ \left(\lambda_{bs}^{(k_{1,1})}, \cdots, \lambda_{bs}^{(k_{I,1})}\right), \cdots, \right. \\ &\qquad \left. \left(\lambda_{bs}^{(k_{1,M})}, \cdots, \lambda_{bs}^{(k_{I,M})}\right) \right\} \\ &\triangleq \left\{\boldsymbol{\lambda}_{bs}^1, \cdots, \boldsymbol{\lambda}_{bs}^M\right\}. \end{aligned}$$

For each element $\boldsymbol{\lambda}_{bs}^i \in \boldsymbol{\Lambda}_{bs}^I, i = 1, \cdots, M$, the objective function in problem (9) becomes $\boldsymbol{\lambda}_{bs}^i\mathbf{r} + q_i$ where q_i is a constant, and we solve the convex optimization problem

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \boldsymbol{\lambda}_{bs}^i\mathbf{r} + q_i \\ \text{subject to} & \mathbf{r} \in \mathcal{C}\left(\bar{\mathbf{P}}\right) \\ & \mathbf{r} \in \boldsymbol{\Delta}_{bs}^I \end{array} \tag{12}$$

in the following way. We first relax the problem to

$$\begin{array}{ll} \underset{\mathbf{r}}{\text{maximize}} & \boldsymbol{\lambda}_{bs}^i\mathbf{r} + q_i \\ \text{subject to} & \mathbf{r} \in \mathcal{C}\left(\bar{\mathbf{P}}\right). \end{array} \tag{13}$$

This is problem (11) with $\boldsymbol{\mu} = \boldsymbol{\lambda}_{bs}^i$ and a constant difference in the objective function, and thus Tse-Hanly solution can be applied to obtain the solution, which we denote as r_{TH}^i. Then, we need to make a binary decision on r_{TH}^i: if $r_{TH}^i \in \boldsymbol{\Delta}_{bs}^i$, r_{TH}^i is the optimal solution to problem (12); otherwise the solution to problem (12) is at the intersection of the boundary of $\boldsymbol{\Delta}_{bs}^i$ and the boundary surface of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$. Again the corresponding rates can be calculated by Tse-Hanly solution and we can obtain the optimal solution to problem (12) by simply comparing the objective function values of these intersection points and choosing the maximum one.

We denote the resulting optimal solution set $\mathfrak{R}_{cand}^* \triangleq \{\mathbf{r}_i^*, i = 1, \cdots, M\}$ where $\mathbf{r}_i^*$ is the solution to problem (12) with $\boldsymbol{\lambda}_{bs}^i$. Then, if we are able to determine that the optimal solution to problem (9) must be in the set $\mathfrak{R}_{cand}^*$, we can claim we have solved problem (9). The optimality, existence and uniqueness of such solution is justified in Theorem 3.

Theorem 3: Given $\boldsymbol{\Delta}_{bs}^I$ and $\boldsymbol{\Lambda}_{bs}^I$, for each element $\boldsymbol{\lambda}_{bs}^i \in \boldsymbol{\Lambda}_{bs}^I, i = 1, \cdots, M$, we solve problem (12), and denote the resulting solution set as $\mathfrak{R}_{cand}^* = \{\mathbf{r}_i^*, i = 1, \cdots, M\}$. Then

a) problem (9) with $Q_i(r_i)$ defined in (6) has one and only one solution;
b) the solution $\mathbf{r}^*$ to problem (9) is in $\mathfrak{R}_{cand}^*$.

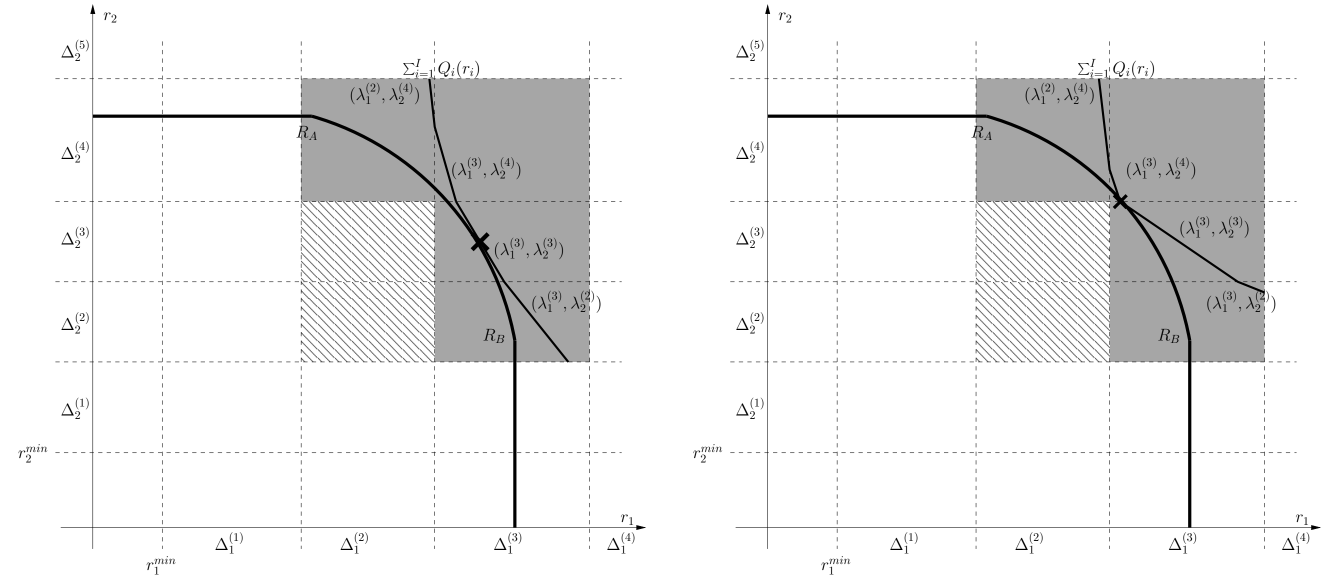
Furthermore, $\mathbf{r}^* = \operatorname{argmax}_{\mathbf{r} \in \mathfrak{R}_{cand}^*} \sum_{i=1}^I Q_i(r_i)$.

Proof: See Appendix B. ■

This method can be further improved in terms of efficiency. From the proof of Theorem 3 we know that if $r_{TH}^i \in \boldsymbol{\Delta}_{bs}^i$, r_{TH}^i is not only the optimal solution to problem (12), but also the optimal solution to problem (9). Thus, an improved scheme to obtain the optimal solution is to first check whether $r_{TH}^i \in \boldsymbol{\Delta}_{bs}^i$ for all possible i's. Only if such r_{TH}^i does not exist for any $i \in \{1, \cdots, M\}$ do we proceed to calculate the intersection points. This will improve the efficiency by avoiding unnecessary calculation of the intersection points.

D. Low-complexity Construction of the Candidate Subset

Now let us go back to the problem of how to obtain the subset $\boldsymbol{\Delta}_{bs}^I$. One natural method is to first obtain the entire



(a) Case 1

(b) Case 2

Fig. 4. Geometric illustrations of how to partition the capacity region $\mathcal{C}\left(\bar{\mathbf{P}}\right)$ and where the optimal solution to problem (9) may be for a two-user case. The gray area is $\boldsymbol{\Delta}_{bs}^I$, and the set of all gray and shadowed rectangles is $\boldsymbol{\Delta}_{cand}^I$. The optimal solution has two possible formats: (a) $\mathbf{r}^* = \mathbf{r}_i^* \in \boldsymbol{\Delta}_i$, or (b) no $\mathbf{r}_i^*$ satisfies $\mathbf{r}_i^* \in \boldsymbol{\Delta}_i$, and the optimal solution is at one intersection of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$ and $\boldsymbol{\Delta}_i$.

boundary surface using Tse-Hanly solution, and then get $\boldsymbol{\Delta}_{bs}^I$ by checking which rate hypercube each point on the boundary surface is in. Unfortunately this method is infeasible, since there are infinitely many points on the boundary surface. To avoid this problem and obtain a low-complexity solution, we propose to use the *candidate set* $\boldsymbol{\Delta}_{cand}^I$, which is defined as the set of all rate hypercubes that can *possibly* have boundary surface points in them. In other words, rate hypercubes that are not in $\boldsymbol{\Delta}_{cand}^I$ will never contain any boundary surface points. We then show using $\boldsymbol{\Delta}_{cand}^I$ instead of $\boldsymbol{\Delta}_{bs}^I$ in the previous algorithm will not change the result, and provide a simple method to construct $\boldsymbol{\Delta}_{cand}^I$.

Since $\boldsymbol{\Delta}_{cand}^I$ has all possible rate hypercubes, the optimal solution is certainly also included, and the previous algorithm can be used to find this solution. We only need to show that in the complement set $\bar{\boldsymbol{\Delta}}_{bs}^I \triangleq \boldsymbol{\Delta}_{cand}^I \backslash \boldsymbol{\Delta}_{bs}^I$, there is no such $\mathbf{r}_i^*$ which is both the solution to problem (11) with the corresponding $\boldsymbol{\mu} = \bar{\boldsymbol{\lambda}}_{bs}^i$ and satisfying $\mathbf{r}_i^* \in \bar{\boldsymbol{\Delta}}_{bs}^i$. This is easy: assume $\mathbf{r}_i^*$ is a solution problem (11), and then according to Tse and Hanly's result, this $\mathbf{r}_i^*$ must be in the boundary surface. Thus it can not be in $\bar{\boldsymbol{\Delta}}_{bs}^i$, which does not have any boundary surface points.

We now show that knowing all the $N = I!$ extreme points is sufficient to construct $\boldsymbol{\Delta}_{cand}^I$. Denote the set of rate hypercubes containing the N extreme points as

$$\begin{aligned}\boldsymbol{\Delta}_{ep}^I &= \left\{ \left(\Delta_{ep}^{(k_{1,1})}, \cdots, \Delta_{ep}^{(k_{I,1})}\right), \cdots, \right. \\ &\qquad \left. \left(\Delta_{ep}^{(k_{1,N})}, \cdots, \Delta_{ep}^{(k_{I,N})}\right) \right\} \\ &\triangleq \left\{ \boldsymbol{\Delta}_{ep}^1, \cdots, \boldsymbol{\Delta}_{ep}^N \right\}\end{aligned}$$

with the corresponding slope set

$$\begin{aligned}\boldsymbol{\Lambda}_{ep}^I &= \left\{ \left(\lambda_{ep}^{(k_{1,1})}, \cdots, \lambda_{ep}^{(k_{I,1})}\right), \cdots, \right. \\ &\qquad \left. \left(\lambda_{ep}^{(k_{1,N})}, \cdots, \lambda_{ep}^{(k_{I,N})}\right) \right\} \\ &\triangleq \left\{ \boldsymbol{\lambda}_{ep}^1, \cdots, \boldsymbol{\lambda}_{ep}^N \right\}.\end{aligned}$$

We further argue that $\boldsymbol{\Delta}_{cand}^I$ can be[3]

$$\begin{aligned}\boldsymbol{\Delta}_{cand}^I = \Bigg\{ \quad & (\Delta_1, \cdots, \Delta_I) \,|\Delta_i \in \left\{\Delta_i^{(1)}, \cdots, \Delta_i^{(N_i)}\right\}, \\ & \Delta_i \geq \min\{\Delta_{ep}^{(k_{i,1})}, \cdots, \Delta_{ep}^{(k_{i,N})}\}, \\ & \Delta_i \leq \max\{\Delta_{ep}^{(k_{i,1})}, \cdots, \Delta_{ep}^{(k_{i,N})}\}, \forall i \Bigg\}.\end{aligned} \tag{14}$$

In fact, the rate set (14) is the minimum I-dimensional hypercube $\{(\Delta_1, \cdots, \Delta_I)\}$ that covers all the N extreme points. Since the boundary surface is strictly within this I-dimensional hypercube, we reach the desired conclusion. For the two-user example shown in Fig. 4, $\boldsymbol{\Delta}_{cand}^I$ is the set of all gray and shadowed rectangles.

This method of using $\boldsymbol{\Delta}_{cand}^I$ has an advantage of low complexity: only $N = I!$ extreme points need to be calculated, and these points have standard "water-filling" solutions. We do not need to calculate any other points on the boundary surface. It has been proved in [2] that the extreme points are achieved by water-filling combined with successive decoding. In the two-user example in Fig. 4, R_A is achieved by first letting user 2 water-fill over the noise level, and then letting user 1

[3] Mathematically, we denote set $\mathbb{A}$ to be smaller than set $\mathbb{B}$ in the following sense: $a \leq b, \forall a \in \mathbb{A}, \forall b \in \mathbb{B}$. Similarly, we define the minimum of N sets $\mathbb{A}_{min} \triangleq \min\{\mathbb{A}_1, \cdots, \mathbb{A}_N\}$ in the sense that $a \leq b, \forall a \in \mathbb{A}_{min}, \forall b \in \mathbb{A}_i, \mathbb{A}_i \neq \mathbb{A}_{min}$. These definitions are valid only if the inequality relationship exists.

water-fill over the sum of noise and interference from user 2. R_B is achieved in the reversed order.

Algorithm 3 I-User Optimal Resource Allocation Algorithm for Line-segment Q-R Models

```
Input: Average power constraint $\bar{\mathbf{P}}$; User $i$'s Q-R model (6)
    with $\left\{\lambda_i^{(1)}, \cdots, \lambda_i^{(N_i)}\right\}$ and $\left\{\Delta_i^{(1)}, \cdots, \Delta_i^{(N_i)}\right\}$, $i = 1, \cdots, I$.
Initialization: Calculate the $I!$ extreme points of $C\left(\bar{\mathbf{P}}\right)$ using
    the standard water-filling solution; Construct $\mathbf{\Delta}_{cand}^{I} = \{\mathbf{\Delta}_i | i = 1, \cdots, M\}$ from (14) with $\mathbf{\Lambda}_{cand}^{I} = \{\boldsymbol{\lambda}_i | i = 1, \cdots, M\}$; $flag = 0$.
Repeat:
  for $i = 1$ to $M$ do
    Solve problem (13) by using Tse-Hanly solution with
    $\boldsymbol{\lambda}_i \in \mathbf{\Lambda}_{cand}^{I}$, and denote as $(\mathbf{r}_i^*, \mathcal{P}_i^*)$.
    if $\mathbf{r}_i^* \in \mathbf{\Delta}_i$ then
      $\mathbf{r}^* = \mathbf{r}_i^*$, $\mathcal{P}^* = \mathcal{P}_i^*$, $flag = 1$, return.
    end if
  end for
  if $flag == 0$ then
    $\mathfrak{R}_{cand}^* = \{\}$, $\mathfrak{P}_{cand}^* = \{\}$.
    for $i = 1$ to $M$ do
      Calculate the intersections of $\mathcal{C}\left(\bar{\mathbf{P}}\right)$ and $\mathbf{\Delta}_i$ using Tse-
      Hanly solution and denote as $\mathfrak{R}_{int}$.
      Choose $\mathbf{r}_{int}^*$ such that $\mathbf{r}_{int}^* = \operatorname{argmax}_{\mathbf{r} \in \mathfrak{R}_{int}} \sum_{i=1}^{I} Q_i(r_i)$.
      Add $\mathbf{r}_{int}^*$ to $\mathfrak{R}_{cand}^*$, add the corresponding $\mathcal{P}_{int}^*$ to
      $\mathfrak{P}_{cand}^*$.
    end for
    $\mathbf{r}^* = \operatorname{argmax}_{\mathbf{r} \in \mathfrak{R}_{cand}^*} \sum_{i=1}^{I} Q_i(r_i)$, and choose the corre-
    sponding $\mathcal{P}^*$ from $\mathfrak{P}_{cand}^*$.
  end if
Return: $(\mathbf{r}^*, \mathcal{P}^*)$
```

Now we are in the position of giving the complete solution to problem (9) in Algorithm 3. The complexity of this algorithm heavily depends on how many hypercubes the boundary of the capacity region dominates. This in turn depends on how granular the line-segment model is. The more granular the line-segment model, the higher the complexity. In fact, the complexity of Algorithm 3 is random, since at any iteration the optimal point could be found and thus the algorithm could have ended. Generally, the complexity of Algorithm 3 is too high to be implementable. However, we want to emphasize that Algorithm 3 is not developed as an operational solution. Instead it is developed only to calculate the theoretical limit of achievable video quality in wireless multiaccess fading channels. We do not intend to recommend it as a practical algorithm to implement. Due to this reason we do not provide numerical examples in this section.

One limitation of Algorithm 3 is that it is not universal for other Q-R models. In this case, one has to resort to the convex optimization theory to obtain numerical solutions [13]. However, since line-segment function is always a good approximation to a continuous function, our algorithm can provide the solution which is very close to the global optimum.

V. Conclusion

This paper shows how resource allocation should be done by a joint consideration of APP-MAC-PHY layers. We demonstrate that the previously known optimal solution becomes suboptimal when APP layer video characteristics are considered. We derive optimal resource allocation policies for different fading channel models from an information-theoretic perspective, and develop efficient algorithms to implement them. Simulation results are shown to support our argument.

Our proposed solution can be viewed as a general theoretic framework for resource allocation. The key ingredient of this framework is to study the resource allocation by jointly considering the upper-layer utility functions (represented by the video quality-rate models) and the lower-layer information-theoretic capacity constraints (represented by the multiaccess capacity region). For example, the solutions developed in this paper, although derived using video quality as the APP layer target, can be extended to other APP layer utility models as long as they have similar properties as the video Q-R models, i.e., any utility-rate function that is monotonically increasing and concave. At the same time, other multi-user capacity regions can be considered in this general framework, e.g., a downlink environment [3].

Appendix A
Proof of Theorem 2

Denote the solution obtained from Algorithm 2 as $\mathbf{r}^* = (r_1, \cdots, r_I)$, and the value of objective function as Q_1^*. Let us assume that the optimal solution to problem (8) is $\mathbf{t}^* = (t_1, \cdots, t_I) \neq \mathbf{r}^*$, and the optimal value of objective function is Q_2^* where $Q_2^* > Q_1^*$. Without loss of generality, we assume $r_1 > t_1$, and denote $\delta_1 = r_1 - t_1$. We define $\Delta F_i\left(r_a, r_b\right) \triangleq \int_{r_a}^{r_b} f_i(z) \mathrm{d}z$.

According to Theorem 1, the optimal solution $\mathbf{t}^*$ satisfies $\sum_{i=1}^{I} t_i = C_{MAX}(\{1, \cdots, I\})$. The solution resulting from Algorithm 2 also satisfies $\sum_{i=1}^{I} r_i = C_{MAX}(\{1, \cdots, I\}) = \sum_{i=1}^{I} t_i$. Thus $\sum_{i=2}^{I} t_i = \sum_{i=2}^{I} r_i + \delta_1$, and we can denote $t_i = r_i + \alpha_i \delta_1, i = 2, \cdots, I, \sum_{i=2}^{I} \alpha_i = 1$. We can then connect Q_1^* with Q_2^* by

$$\begin{aligned} Q_2^* &= Q_1^* + \Delta F_1\left(r_1, r_1 - \delta_1\right) + \sum_{i=2}^{I} \Delta F_i\left(r_i, r_i + \alpha_i \delta_i\right) \\ &> Q_1^*. \end{aligned} \tag{15}$$

This suggests

$$\Delta F_1\left(r_1, r_1 - \delta_1\right) + \sum_{i=2}^{I} \Delta F_i\left(r_i, r_i + \alpha_i \delta_i\right) > 0. \tag{16}$$

However, inequality (16) contradicts with the principle of Algorithm 2. Rates are always allocated to users for which they lead to the maximum area increase in Algorithm 2. Inequality (16), on the other hand, says that giving a total rate δ_1 to user $2, \cdots, I$ will result in larger area $\sum_{i=2}^{I} \Delta F_i\left(r_i, r_i + \alpha_i \delta_i\right)$ than giving it to user 1: $\sum_{i=2}^{I} \Delta F_i\left(r_i, r_i + \alpha_i \delta_i\right) > -F_1\left(r_1, r_1 - \delta_1\right)$. From this contradiction, we conclude that such $\mathbf{t}^* = (t_1, \cdots, t_I) \neq \mathbf{r}^*$ with $Q_2^* > Q_1^*$ does not exist, and $\mathbf{r}^* = (r_1, \cdots, r_I)$ is the optimal solution.

Appendix B
Proof of Theorem 3

The existence of the optimal solution comes directly from that the set $C\left(\bar{\mathbf{P}}\right)$ is closed and bounded, and $Q(r)$ is finite for any finite r. To prove the uniqueness of the solution, we first prove the following lemma:

Lemma 2: Define $f\left(\mathbf{r}\right) = \sum_{i=1}^{I} Q_i\left(r_i\right)$ where $Q_i\left(r_i\right)$ is defined in (6). If $f\left(\mathbf{r}\right) = f\left(\mathbf{s}\right) = C$ and $\mathbf{r} \neq \mathbf{s}$ where C is a constant, then $f\left(\alpha\mathbf{r} + (1-\alpha)\,\mathbf{s}\right) \geq C$ for any $0 \leq \alpha \leq 1$.

Proof: For any $0 \leq \alpha \leq 1$, assume $\mathbf{r} \in \left(\Delta_1^{(k_1)}, \cdots, \Delta_I^{(k_I)}\right)$, $\mathbf{s} \in \left(\Delta_1^{(l_1)}, \cdots, \Delta_I^{(l_I)}\right)$ and $\alpha\mathbf{r} + (1-\alpha)\,\mathbf{s} \in \left(\Delta_1^{(h_1)}, \cdots, \Delta_I^{(h_I)}\right)$. We have $\Delta_1^{(k_i)} \leq \Delta_1^{(h_i)} \leq \Delta_1^{(l_i)}, \forall i = 1, \cdots, I$. Since $Q_i(r_i)$ is a concave function, we have $Q_i\left(\alpha r_i + (1-\alpha)s_i\right) \geq \alpha Q_i\left(r_i\right) + (1-\alpha)Q_i\left(s_i\right), r_i \in \Delta_1^{(k_i)}, s_i \in \Delta_1^{(k_i)}, \alpha r_i + (1-\alpha)s_i \in \Delta_1^{(h_i)}$. Thus

$$\begin{aligned} f\left(\alpha\mathbf{r} + (1-\alpha)\,\mathbf{s}\right) &= \sum_{i=1}^{I} Q_i\left(\alpha r_i + (1-\alpha)s_i\right) \\ &\geq \sum_{i=1}^{I} \alpha Q_i\left(r_i\right) + (1-\alpha)Q_i\left(s_i\right) \\ &= C \end{aligned}$$

■

Now with the help of Lemma 2, we can prove there is only one solution to problem (9). Let us assume there are two optimal solutions, $\mathbf{r}_1$ and $\mathbf{r}_2$, such that $Q^* = f\left(\mathbf{r}_1\right) = f\left(\mathbf{r}_2\right)$. Consider $\alpha\mathbf{r}_1 + (1-\alpha)\mathbf{r}_2$, due to the convexity of $C\left(\bar{\mathbf{P}}\right)$ we have $\alpha\mathbf{r}_1 + (1-\alpha)\mathbf{r}_2 \in C\left(\bar{\mathbf{P}}\right)$. From Lemma 2 we know $f\left(\alpha\mathbf{r}_1 + (1-\alpha)\mathbf{r}_2\right) \geq Q^*$. This contradicts the fact that $\mathbf{r}_1$ and $\mathbf{r}_2$ are optimal solutions, because there is no linear part at the boundary surface of $C\left(\bar{\mathbf{P}}\right)$ [2]. We thus conclude that there is only one optimal solution.

To prove that the optimal solution must be in $\mathfrak{R}^*_{cand}$, we use the fact that any locally optimal solution is also globally optimal for convex optimization problems [13]. From Theorem 1, the optimal solution must be at the boundary surface, and thus must be in $\boldsymbol{\Delta}^I_{bs}$. Let us further assume that the optimal solution to problem (9) satisfies $\mathbf{r}^* \in \boldsymbol{\Delta}^i_{bs}$. In this hypercube, the original problem (9) is equivalent to problem (12). According to the definition of $\mathfrak{R}^*_{cand}$, the locally optimal solution to problem (9) in $\boldsymbol{\Delta}^i_{bs}$ is $\mathbf{r}^*_i \in \mathfrak{R}^*_{cand}$. Since locally optimal solution is also globally optimal, we conclude that $\mathbf{r}^* = \mathbf{r}^*_i \in \mathfrak{R}^*_{cand}$.

Acknowledgment

The authors would like to thank the Associate Editor, Prof. Dong In Kim, and the anonymous reviewers for their comments that greatly improve this paper.

References

[1] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE International Conference on Communications*, vol. 1, June 1995, pp. 331–335.

[2] D. Tse and S. Hanly, "Multiaccess fading channels–part I: polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2796–2815, Nov. 1998.

[3] D. Tse, "Optimal power allocation over parallel gaussian broadcast channels," in *Proc. International Symposium on Information Theory (ISIT)*, 1997, p. 27.

[4] ——, "Multi-user diversity and proportional fairness," Sept. 2002, U.S. Patent 6,449,490.

[5] A. Eryilmaz, R. Srikant, and J. R. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Networking*, vol. 13, no. 2, pp. 411–424, Apr. 2005.

[6] R. Negi and S. Goel, "An information-theoretic approach to queuing in wireless channels with large delay bounds," in *Proc. Globecom 2004*, vol. 1, Nov. 2004, pp. 116–122.

[7] M. van der Schaar, Y. Andreopoulos, and Z. Hu, "Optimized scalable video streaming over IEEE 802.11 a/e HCCA wireless networks under delay constraints," *IEEE Trans. Mobile Comput.*, vol. 5, no. 6, pp. 755–768, June 2006.

[8] F. Zhai, "Cross-layer resource allocation for video transmission over packet lossy networks," Ph.D. dissertation, Northwestern University, Evanston, IL, 2004.

[9] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley Interscience, 2006.

[10] S. Shamai and A. D. Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading channels–part I," *IEEE Trans. Inform. Theory*, vol. 43, no. 6, pp. 1877–1894, Nov. 1997.

[11] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Mag.*, vol. 15, pp. 23–50, Nov. 1998.

[12] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 1012–1032, June 2000.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge University Press, 2004.

[14] B. Rimoldi and R. Urbanke, "A rate-splitting approach to the Gaussian multiple-access channel," *IEEE Trans. Inform. Theory*, vol. 42, no. 2, pp. 364–375, Mar. 1996.

[15] M. van der Schaar and P. Chou, Eds., *Multimedia over IP and Wireless Networks: Compression, Networking, and Systems*. Elsevier, Mar. 2007.

[16] T. Stockhammer and M. Bystrom, "H.264/AVC data partitioning for mobile video communication," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 545–548, Oct. 2004.

[17] M. van der Schaar and D. Turaga, "Cross-layer packetization and retransmission strategies for delay-sensitive wireless multimedia transmission," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 185–197, Jan. 2007.

Cong Shen (S'01) received the B.S. and M.S. degrees in 2002 and 2004, respectively, from the Department of Electronic Engineering, Tsinghua University, Beijing, China. He is currently working toward the Ph.D. degree in the Electrical Engineering Department, University of California, Los Angeles (UCLA). His research interest is on general communication theory with emphasis on wireless communications.

Mihaela van der Schaar (SM'04) received the Ph.D. degree from Eindhoven University of Technology, the Netherlands, in 2001. She is currently an associate professor in the Electrical Engineering Department at UCLA. She received the NSF CAREER Award in 2004, the IBM Faculty Award in 2005 and 2007, the Okawa Foundation Award in 2006, the Best IEEE Transactions on Circuits and Systems for Video Technology Paper Award in 2005, and the Most Cited Paper Award from Signal Processing: Image Communication between 2004 and 2006.