Rate Adaptation and Admission Control for Video Transmission With Subjective Quality Constraints

Chao Chen, *Member, IEEE*, Xiaoqing Zhu, *Member, IEEE*, Gustavo de Veciana, *Fellow, IEEE*, Alan C. Bovik, *Fellow, IEEE*, and Robert W. Heath, Jr., *Fellow, IEEE*

***Abstract*—Adapting video data rate during streaming can effectively reduce the risk of playback interruptions caused by channel throughput fluctuations. The variations in rate, however, also introduce video quality fluctuations and thus potentially affects viewers' Quality of Experience (QoE). We show how the QoE of video users can be improved by rate adaptation and admission control. We conducted a subjective study wherein we found that viewers' QoE was strongly correlated with the empirical cumulative distribution function (eCDF) of the predicted video quality. Based on this observation, we propose a rate-adaptation algorithm that can incorporate QoE constraints on the empirical cumulative quality distribution per user. We then propose a threshold-based admission control policy to block users whose empirical cumulative quality distribution is not likely to satisfy their QoE constraint. We further devise an online adaptation algorithm to automatically optimize the threshold. Extensive simulation results show that the proposed scheme can reduce network resource consumption by 40% over conventional average-quality maximized rate-adaptation algorithms.**

***Index Terms*—Quality of experience, video transport, rate adaptation, admission control, wireless networks.**

I. Introduction

VIDEO traffic is currently a rapidly growing fraction of the data traffic on wireless networks. As reported in [3], video traffic accounted for more than 50% of the mobile data traffic in 2012. Efficiently utilizing network resources to satisfy video users' expectations regarding their Quality of Experience (QoE) is an important research topic. In this paper, we study approaches to share wireless down-link resources among video users via QoE-based rate adaptation and admission control.

We focus on a setting in which stored video content is streamed over wireless networks. When a video is streamed, the received video data is first buffered at the receiver and then played out to the viewer. Because the throughput of a wireless channel generally varies over time, the amount of buffered video decreases when the channel throughput falls below the current video data rate. Once all the video data buffered at the receiver has been played out, the playback process stalls, significantly affecting the viewers QoE [4]. To address this problem, various video rate-adaptation techniques based on scalable video coding or adaptive bitrate switching have been proposed to match the video data rate to the varying channel capacity [5]–[9]. Although these rate adaptation techniques can effectively reduce the risk of playback interruptions, the variable bitrate causes quality fluctuations, which, in turn, affect viewers' QoE. In most existing rate-adaptation algorithms such as [10]–[15], the average video quality is employed as the proxy for QoE. The average quality, however, does not reflect the impact of quality fluctuations on the QoE, i.e., two videos with the same average quality can have significantly different levels of quality fluctuation. In this paper, we propose to characterize and predict the users' QoE using the second order empirical cumulative distribution function (2^{nd}-order eCDF) of the delivered video quality; this is defined as

$$\mathrm{F}^{(2)}(x;\mathrm{q}) = \frac{1}{T}\sum_{t=1}^{T}\max\{x-\mathrm{q}(t),0\}, \tag{1}$$

where $\mathrm{q}(t)$ represents the predicted quality [16] of the t^{th} second of the video and T is the video length. Note that $\max\{x-\mathrm{q}(t),0\} > 0$ if and only if $\mathrm{q}(t) < x$, so the 2^{nd}-order eCDF captures for how long and by how much the predicted video quality falls below x. If we interpret x as the threshold below which users judge the video quality to be unacceptable, then the 2^{nd}-order eCDF reflects the impact of the unacceptable periods on the QoE. Since it has been recognized that the worst parts of a video tend to dominate the overall quality of an entire video [17]–[21], the 2^{nd}-order eCDF can be used to predict the QoE.

The efficacy of the 2^{nd}-order eCDF in capturing QoE can be validated through subjective experiments. In [1] and [2], we reported a subjective study of this type. For each of the 15 quality-varying videos involved in the subjective study, we asked 25 subjects to score its quality. We computed the linear correlation coefficients (LCCs) between various QoE metrics and the subjects' Mean Opinion Scores (MOSs) in Table I. The 2^{nd}-order eCDF achieves the strongest linear correlation (0.84). In comparison, the average video quality only achieves a correlation of 0.57. This lends strong support for eCDF as a good proxy for video QoE.

Manuscript received November 18, 2013; revised March 29, 2014; accepted June 12, 2014. Date of publication July 08, 2014; date of current version January 20, 2015. This work was supported in part by Intel Inc. and Cisco Corp. under the VAWN program. The details of the subjective study mentioned in this paper have been presented in [1] and [2]. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Béatrice Pesquet-Popescu.

C. Chen, G. de Veciana, A. Bovik, and R. Heath are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712-0240 USA (email: tochenchao@gmail.com; gustavo@ece.utexas.edu; bovik@ece.utexas.edu; rheath@ece.utexas.edu).

Xiaoqing Zhu is with the Advanced Architecture and Research Group, Cisco Systems, San Jose, CA 95134 USA (email: xiaoqzhu@cisco.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2014.2337277

TABLE I
The Linear Correlation Coefficients (LCCs) of Several Metrics With QoE. The Metrics Mean$\{q\}$, $\min\{q\}$ and var$\{q\}$ are the Average Value, the Maximum Value and the Variance of the Time Series $q(1), \ldots, q(T)$, Where $q(t)$ is the Predicted Video Quality at Time t

QoE metrics	mean$\{q\}$	min$\{q\}$	mean$\{q\} + \mu\sqrt{\mathrm{var}\{q\}}$	$F^{(2)}(x; q)$
LCC	0.5659	0.4022	0.7377	0.8446

In this paper, we design adaptive video streaming algorithms that incorporate QoE constraints on the 2^{nd}-order eCDFs of the video qualities seen by users. In particular, we consider a wireless network in which a base station transmits videos to multiple users. The user population is dynamic, i.e., users arrive and depart from the network at random times. When a new user joins the network, the base station starts streaming a video to the user. A rate adaptation algorithm is employed to control the video data rate of all active video streams according to varying wireless channel conditions.

When the base station is shared by too many users simultaneously, the QoE served to each user can be poor. Instead of serving every user with poor QoE, it is preferable to satisfy the QoE constraints of existing users by selectively blocking newcomers. Therefore, in addition to rate adaptation algorithms, we introduce a new admission control strategy that is designed to maximize the number of video users satisfying the QoE constraints on their 2^{nd}-order eCDFs. As will be shown in the paper, although the admission control strategy damages the QoE of the blocked users, the overall percentage of users satisfying the QoE constraints among both admitted and blocked users can be significantly improved. The contributions of our work are twofold:

- *An online rate-adaptation algorithm aimed at meeting QoE constraints based on the 2^{nd}-order eCDFs of users' video qualities.*
  Since the 2^{nd}-order eCDF is determined by the overall spatio-temporal pattern of video, simply maximizing the video quality all the times is not sufficient to satisfy the QoE constraints. Instead, we propose an online rate-adaptation algorithm that jointly considers the channel conditions, the video rate-quality characteristics, and the 2^{nd}-order eCDFs of all video users. We show significant performance gains over conventional average-quality optimized algorithms.
- *An admission control strategy, which blocks video users who will likely be unable to meet the constraints on their 2^{nd}-order eCDFs.*
  Specifically, we propose an algorithm that predicts the video quality experienced by each user who is new to the network. We then employ a threshold-based admission control policy to block those users whose estimated qualities fall below the threshold. An online algorithm is proposed to adjust the threshold to its optimal value. The proposed admission control strategy further improves the performance of our rate-adaptation algorithm, especially when the network resources are limited.

The remainder of this paper is organized as follows: Section II discusses related work. In Section III, we give an overview of the structure of the video streaming system studied in this paper. In Section IV, we explain our system model and the QoE constraints. The proposed rate adaptation algorithm and the admission control strategy are introduced in Sections V and VI. We demonstrate their performance via extensive simulation results. Section VII concludes the paper with discussions on future work.

II. Related Work

Let us first review related work in QoE assessment, rate adaptation, and admission control.

A. QoE Metrics for Rate-Adaptive Video Steaming

Most existing rate-adaptation algorithms such as [10]–[15], [22], [23] employ average video quality as the proxy of QoE due to its simplicity in analysis. As shown in Table I, however, average quality does not efficiently capture the impact of quality fluctuations. To address this problem, temporal quality pooling has been studied [17]–[21]. The pooling algorithms, however, have only been designed and validated for short videos that are a few seconds long. Furthermore, they are too complicated to be incorporated into a tractable analytical framework for the design of rate-adaptation algorithms. The authors of [24] propose a simple QoE metric, which is the weighted sum of the time-average and the standard deviation of the predicted video quality. As shown in Table I, this metric achieves a better correlation of 0.73 than the time-average quality. The QoE metric that we propose here achieves an even higher correlation of 0.84. Moreover, because the 2^{nd}-order eCDF is simply a temporal average of the function $\max\{x - q(t), 0\}$, its analysis is just as simple as for average video quality.

B. Rate Adaptive Video Streaming

Extensive research efforts have been applied to the problem of rate adaptation for wireless video streaming. Most existing work employs the time-average video quality as the QoE metric due to its simplicity [10]–[15], [22], [23]. In [25], a rate adaptation algorithm is proposed to optimize the QoE metric of [24]. Although the proposed algorithm mitigates video quality variations, it assumes a fixed set of users and thus does not incorporate the impact of user arrival and departure on video quality. In [26], the quality fluctuations caused by user arrival and departure are analyzed for large wireless links shared by many video streams. The algorithms proposed here do not rely on such assumptions and can be applied to small wireless cells such as Wi-Fi networks. In [27], a distributed flow control algorithm is presented to achieve the utility max-min fair bandwidth allocation if the slope of the utility function is lower-bounded by a certain positive value. Note that the 2^{nd}-order eCDF is the time-average of the utility function $\max\{x - q(t)\}$, whose minimum slope is 0. Thus the algorithm proposed in [27] is not applicable.

C. Admission Control for Video Streaming

Admission control for variable bitrate videos has been studied in [28]–[30]. In [28], an admission control algorithm is proposed for variable bitrate videos stored on disk arrays and transmitted over cable television networks. In contrast to wireless networks, the transmission bottleneck of cable television networks is the buffer at the disk array. The admission control

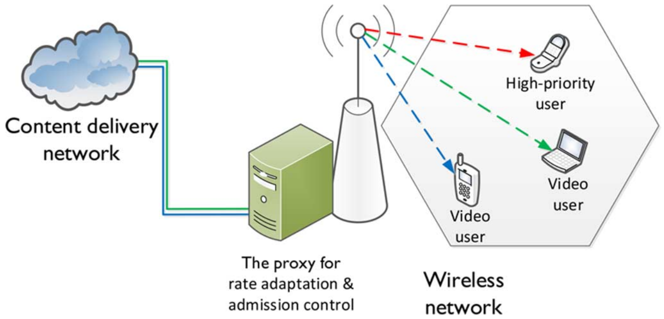


Fig. 1. The wireless network considered here: The video is stored at the content delivery network. The proxy for rate adaptation and admission control is colocated with the base station. The base station serves both video users and high-priority users.

algorithm is designed to ensure that the buffer does not overflow while the video is played back continuously. In [29] and [30], threshold-based admission control algorithms for video streams delivered over packet-switch networks are studied. In [29], thresholds are applied to the number of video users. The heterogeneity in the data rate of the video streams is not considered. In [30], the threshold is applied to the aggregated data rate requested by admitted videos. In both [29] and [30], a statistical model of the video traffic is necessary to optimize the admission threshold. In practice, however, modeling the video traffic is difficult. Our proposed admission control strategy does not rely on the *a priori* knowledge of the traffic statistics. The threshold is learned online and optimized automatically.

III. System Overview

We first discuss the architecture of the wireless networks considered in this paper. Then, we explain how the proposed online rate adaptation algorithm and threshold-based admission control strategy fit into the existing network architecture.

A. Architecture of the Wireless Network

We consider a wireless network where video users share the down-link with high-priority traffic (e.g., voice traffic) and thus video users will need to adapt their rates accordingly. (see Fig. 1). All users arrive to and depart from the network at random times. A high-priority user requires a random amount of wireless resource (e.g., transmission time in TDMA systems) per unit time throughout its sojourn in the system. The transmission resources not used by the high-priority users can be allocated to video users. When a video user arrives, it requests a video that is stored at a content delivery network (CDN) and is streamed to the user via the base station. When a video is being streamed, the video data is first delivered to a receive buffer and then decoded for display. Paralleling prior work such as [15], [25], [31], we assume a proxy is colocated with the base station. We treat the high-priority users as background traffic, and the proxy is only used to control the video streams.

B. Video Streaming Proxy

The function of the proxy is twofold (see Fig. 2). First, when a video user arrives, the proxy decides whether the user should be admitted to share the channel or not. Second, for admitted video users, the proxy adapts the transmission data rates according to the varying channel conditions. The admission control strategy only acts on newly arrived video users, and the rate adaptation algorithm does its best to satisfy QoE constraints for all admitted video users. The rate adaptation algorithm accompanies the admission control algorithm by feeding back necessary information (see Section V-B for more detail) regarding the current status of the admitted users.

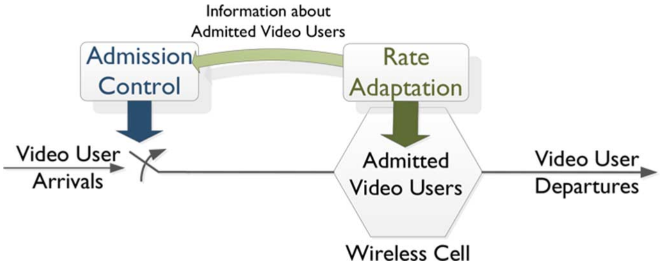


Fig. 2. The proposed QoE-constrained video streaming system: Admission control only affects newly arrived video users and the rate adaptation algorithm does its best to satisfy the QoE constraints of all admitted video users.

We assume the proxy operates in a time-slotted manner where the duration of each slot is ΔT seconds. The admission control and the rate adaptation actions are conducted at the beginning of every time slot. Note that, in a video stream, the video frames are partitioned in Groups of Pictures (GoPs) and the video data rate can only be adapted at the boundary of GoPs [32], [33]. Because the duration of a GoP is usually configured to be larger than one second to achieve high compression efficiency, we assume ΔT is at least 1 second.

IV. System Model

Before proceeding further, we introduce the notation used in the paper. Then, we describe the channel model and the video rate-quality model. At the end of this section, we explain the QoE constraints considered in our problem formulation.

A. Notation

In the rest of the paper, the time slots are indexed with $t = 1, 2, \ldots$. The notation $(\mathsf{x}(t), t = 1, 2 \ldots)$ denote a discrete time series. Lower-case symbols such as a denote scalar variables and boldface symbols such as $\mathbf{a}$ denote vectors. Random variables are denoted by uppercase letters such as A. Calligraphic symbols such as $\mathcal{A}$ denote sets, while $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$. The set of positive integers is denoted $\mathbb{N}^+$. The set of real numbers is denoted $\mathbb{R}$. Finally, the function $[x]^+ = \max\{x, 0\}$.

B. Wireless Channel Model

We label users (including both video users and high-priority users) according to their arrival times, i.e., user u is the u^{th} user to arrive to the network. For each user u, we let A_u and D_u be random variables denoting the arrival and departure times, respectively. The time spent by a user in the network is denoted by $\mathsf{T}_u = \mathsf{D}_u - \mathsf{A}_u + 1$.

We let by $\mathcal{U}^{\mathrm{p}}(t)$ be the set of high-priority users in the network at slot t. For each high-priority user $u \in \mathcal{U}^{\mathrm{p}}(t)$, we let the random variable $\mathsf{W}_u(t)$ represent the amount of data received in slot t. The data rate is thus $\mathsf{R}_u(t) = \mathsf{W}_u(t)/\Delta T$. We denote by $\mathcal{U}^{\mathrm{v}}(t)$ the set of video users that would be in the network at slot t if all were admitted. The set of video users in $\mathcal{U}^{\mathrm{v}}(t)$ who are actually admitted to share the wireless channel is denoted by $\mathcal{U}^{\mathrm{av}}(t)$. We assume that an admission decision is made upon the arrival of each video user. Once admitted, the video user shares the channel until it leaves the network. For an admitted video user $u \in \mathcal{U}^{\mathrm{av}}(t)$, we denote by $\mathrm{w}_u(t)$ the amount of received video data in slot t. The video data rate delivered to the user is thus $\mathrm{r}_u(t) = \mathrm{w}_u(t)/\Delta T$. We call $\mathbf{r}(t) = (\mathrm{r}_u(t) : u \in \mathcal{U}^{\mathrm{av}}(t))$ the *video rate vector* at time t. Because high-priority users are treated as background traffic, the proxy only controls the video rate vector and regards $(\mathsf{R}_u(t) : u \in \mathcal{U}^{\mathrm{p}}(t))$ as exogenous variables.

We assume the set of video rate vectors that can be supported by the wireless channel is given by

$$\mathcal{C}(t) = \{\mathbf{r} : \mathrm{C}_t(\mathbf{r}) \leq 0\}, \tag{2}$$

where $\mathrm{C}_t : \mathbb{R}^{|\mathcal{U}^{\mathrm{av}}(t)|} \rightarrow \mathbb{R}$ is a time-varying multivariate convex function. The specific form of $\mathrm{C}_t(\,\cdot\,)$ depends on the multiuser multiplexing techniques used in the wireless network. For example, in a time-division multiple access (TDMA) system [15], [25], the channel is occupied by a single user at any moment. Denote by $\mathsf{P}_u(t)$ the peak transmission rate of user u, i.e., the transmission rate at which user u can be served during slot t. Then, in slot t, video user $u \in \mathcal{U}^{\mathrm{av}}(t)$ spends $\frac{\mathrm{w}_u(t)}{\mathsf{P}_u(t)}$ seconds to download video. Similarly, each high-priority user $u' \in \mathcal{U}^{\mathrm{p}}(t)$ expends $\frac{\mathsf{W}_{u'}(t)}{\mathsf{P}_{u'}(t)}$ seconds downloading data. Since the total transmission across all users is less than ΔT, we have $\sum_{u\in\mathcal{U}^{\mathrm{av}}(t)} \frac{\mathrm{w}_u(t)}{\mathsf{P}_u(t)} + \sum_{u'\in\mathcal{U}^{\mathrm{p}}(t)} \frac{\mathsf{W}_{u'}(t)}{\mathsf{P}_{u'}(t)} \leq \Delta T$. Because $\mathrm{r}_u(t) = \mathrm{w}_u(t)/\Delta T$ and $\mathsf{R}_{u'}(t) = \mathsf{W}_{u'}(t)/\Delta T$, we have $\sum_{u\in\mathcal{U}^{\mathrm{av}}(t)} \frac{\mathrm{r}_u(t)}{\mathsf{P}_u(t)} + \sum_{u'\in\mathcal{U}^{\mathrm{p}}(t)} \frac{\mathsf{R}_{u'}(t)}{\mathsf{P}_{u'}(t)} \leq 1$. Therefore, for TDMA systems, we have $\mathrm{C}_t(\mathbf{r}(t)) = \sum_{u\in\mathcal{U}^{\mathrm{av}}(t)} \frac{\mathrm{r}_u(t)}{\mathsf{P}_u(t)} + \sum_{u'\in\mathcal{U}^{\mathrm{p}}(t)} \frac{\mathsf{R}_{u'}(t)}{\mathsf{P}_{u'}(t)} - 1$.

In rate-adaptive video streaming systems such as [7]–[9], the video data rate can only take values in a finite and discrete set. In our problem formulation, we relax the constraint and allow $\mathrm{r}_u(t)$ to take values in a compact interval $[\mathrm{r}_u^{\min}(t), \mathrm{r}_u^{\max}(t)]$, where $\mathrm{r}_u^{\min}(t)$ and $\mathrm{r}_u^{\max}(t)$ denote the minimum and maximum data rate available for user u, respectively. Letting $\mathcal{R}(t) = \Pi_{u\in\mathcal{U}^{\mathrm{av}}(t)}[\mathrm{r}_u^{\min}(t), \mathrm{r}_u^{\max}(t)]$, we have

$$\mathbf{r}(t) \in \mathcal{R}(t). \tag{3}$$

In our algorithm implementation, we round up the optimal video data rate obtained under this relaxed constraint to the nearest available data rate.

C. Video Rate-Quality Model

We assume that the quality of the video downloaded in each slot is represented by a Difference Mean Opinion Score (DMOS) [34], which ranges from 0 to 100 where lower values indicate better quality. To represent video quality more naturally, so that higher numbers indicate better video quality, we deploy a Reversed DMOS (RDMOS). Denote by $\mathrm{q}_u^{\mathrm{dmos}}(t)$ the DMOS of the video delivered to user u at slot t, the RDMOS is given by $\mathrm{q}_u^{\mathrm{rdmos}}(t) = 100 - \mathrm{q}_u^{\mathrm{dmos}}(t)$. We employ the following rate-quality model to predict $\mathrm{q}_u^{\mathrm{rdmos}}(t)$ using the video data rate $\mathrm{r}_u(t)$:

$$\mathrm{q}_u(t) = \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t), \tag{4}$$

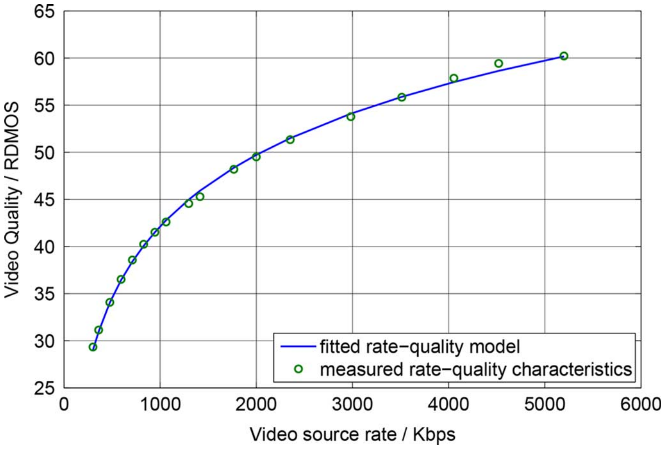


Fig. 3. The performance of the rate-quality model on one second of a video randomly chosen from the database [1]. The rate-quality characteristics are shown in circles while the fitted rate-quality model (4) is the solid line.

where $\mathrm{q}_u(t)$ is the predicted RDMOS. The model parameters $\alpha_u(t)$ and $\beta_u(t)$ can be determined by minimizing the prediction error between $\mathrm{q}_u^{\mathrm{rdmos}}(t)$ and $\mathrm{q}_u(t)$. For stored video streaming, the video file is stored at the CDN. Thus, the model parameters in (4) can be obtained before transmission. Here, we assume the parameters $\alpha_u(t)$ and $\beta_u(t)$ are known *a priori*.

We validated this model on a video database that includes twenty-five different pristine and representative videos [1]. The rest of the database is created by encoding and decoding each video at different rates with the widely used H.264 codec [35]. Then, we predict the RDMOSs for each second of the decoded videos using the high-accuracy ST-RRED (Spatial-Temporal Reduced Reference Entropic Difference) index [36]. In Fig. 3, we show the rate-RDMOS mapping of one second of a video randomly chosen from the database. It may be observed that the model (4) can accurately predict the RDMOSs. On the whole database, the mean prediction error of (4) is less than 1.5, which is visually negligible. Thus, in the following, we shall refer to $\mathrm{q}_u(t)$ as the video quality.

D. Constraints on the Quality of Experience

We capture video users' QoE using the 2^{th}-order eCDF $\mathrm{F}^{(2)}(x; \mathrm{q})$, which was defined in (1). As illustrated in Fig. 4(a), for a given x, the right-hand side of (1) is proportional to the area where $\mathrm{q}(t)$ falls below x. If $\mathrm{q}(t)$ falls below x for a long while, the QoE is poor and $\mathrm{F}^{(2)}(x; \mathrm{q})$ is large. Otherwise, the QoE is good and $\mathrm{F}^{(2)}(x; \mathrm{q})$ is small.

To justify the use of the $\mathrm{F}^{(2)}(x; \mathrm{q})$ as the QoE metric, we conducted a subjective study following the guidelines of [34]. The study involved twenty-five subjects and fifteen quality-varying long videos (for more details, see [1] and [2]). Based on the subjects' feedback, we obtained the Mean Opinion Scores (MOSs) of each video's overall quality. Given an x, we computed $\mathrm{F}^{(2)}(x; \mathrm{q})$ for all the videos in the database and then calculated the linear correlation coefficient (LCC) between the computed $\mathrm{F}^{(2)}(x; \mathrm{q})$s and the MOSs. In Fig. 4(b), we plot the absolute value of the LCCs as a function of x. We found that, at $x^* = 37$, $\mathrm{F}^{(2)}(x^*; \mathrm{q})$ achieves a strong correlation of 0.84 with the MOSs. Since $\mathrm{F}^{(2)}(x^*; \mathrm{q})$ is determined by the area where $\mathrm{q}(t)$ falls below x^*, we interpret x^* as the users' *video*

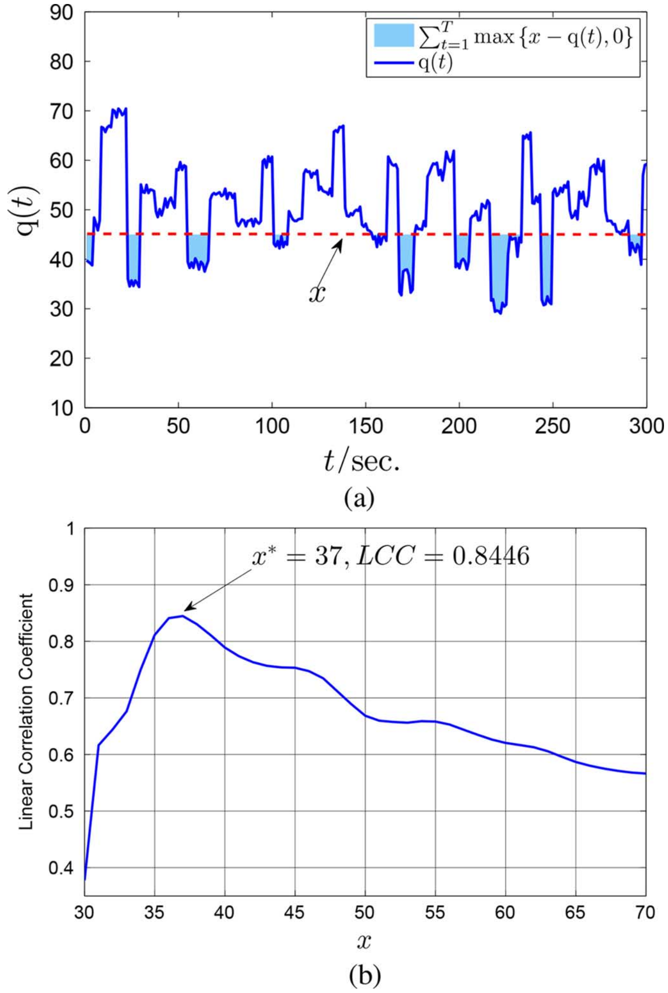


Fig. 4. (a) An example of $\mathrm{F}^{(2)}(x;\mathrm{q})$ at $x=45$; (b) The absolute value of the linear correlation coefficient (LCC) between $\mathrm{F}^{(2)}(x;\mathrm{q})$ and the mean opinion scores.

quality expectation, which is used by the users as a threshold in judging whether the video quality is acceptable or not. In our subjective study, all subjects viewed the videos in a controlled environment and every subject viewed the videos on the same device. Broadly speaking, the video quality expectation x^* can be environment-dependent. For example, viewers tend to have higher expectation for videos shown on a laptop than videos shown on a smartphone. Therefore, in a practical wireless network, x^* may vary across users. In the following, we denote by x_u^* the video quality expectation of user u and study the following two cases:

Case I: Users' video quality expectation is unavailable. If user u is in the system from A_u to D_u and sees video qualities $(\mathrm{q}_u(t): t\in[\mathsf{A}_u,\mathsf{D}_u])$, according to (1), its 2nd-order eCDF is given by

$$\mathrm{F}^{(2)}(x;\mathrm{q}_u)=\frac{1}{\mathsf{T}_u}\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u}[x-\mathrm{q}_u(t)]^+. \tag{5}$$

If the users' video quality expectation x_u^* is not known *a priori*, we may impose constraints on all x and for all video users. In particular, we consider the following QoE constraints:

$$\mathrm{F}^{(2)}(x;\mathrm{q}_u)\le \mathrm{h}(x), \forall x\in[0,100], \forall u\in\cup_{t=1}^{\infty}\mathcal{U}^{\mathrm{av}}(t), \tag{6}$$

TABLE II
NOTATION SUMMARY

$\mathsf{A}_u, \mathsf{D}_u$	The arrival and departure time of user u
T_u	Time spent by user u in the network
$\mathcal{U}^{\mathrm{v}}(t)$	Video users at time t;
$\mathcal{U}^{\mathrm{av}}(t)$	Admitted video users at time t;
$\mathrm{r}_u(t)$	Data rate of user u at t
$\mathcal{C}(t)$	Rate region at t
$\mathcal{R}(t)$	The set of available source rates at t
$\mathrm{q}_u(t)$	Video quality of user u at t
$\mathrm{F}^{(2)}(x;\mathrm{q}_u)$	The second order eCDF of user u
x_u^*	Video quality expectation of user u
$\mathcal{I}$	Constrained points on eCDFs
$\mathrm{h}(x_i)$	The QoE constraint at $x_i\in\mathcal{I}$
$\mathcal{G}$	Quality expectations of video users
$\mathcal{U}_j^{\mathrm{av}}$	Admitted video users with $x_u^*=g_j\in\mathcal{G}$
h_j	The QoE constraint for the users in $\mathcal{U}_j^{\mathrm{av}}$

where $\mathrm{h}(x)$ is a function of x. In practice, we cannot apply constraints on all values of $x\in[0,100]$. Therefore, we consider a relaxed version of (6) as follows:

$$\mathrm{F}^{(2)}(x_i;\mathrm{q}_u)\le \mathrm{h}(x_i), \forall x_i\in\mathcal{I}, \forall u\in\cup_{t=1}^{\infty}\mathcal{U}^{\mathrm{av}}(t). \tag{7}$$

Here, $\mathcal{I}$ is a discrete set of points on $[0,100]$. The following property of 2nd-order eCDFs shows that (7) will approximate (6) if $\mathcal{I}$ is dense. Its proof is given in Appendix A.

Theorem 1: Let $\bar{\mathrm{h}}(x)$ be the piece-wise linear function that connects the points $\{(x_i,\mathrm{h}(x_i)):\forall x_i\in\mathcal{I}\}$. The constraint (7) is equivalent to $\mathrm{F}^{(2)}(x;\mathrm{q}_u)\le\bar{\mathrm{h}}(x), \forall x\in[0,100]$.

Case II: Users' video quality expectation is known. We also consider the case where users' video quality expectation x_u^* is known or specified by the service provider. For example, users with different viewing devices tend to have different quality expectations. Desktop users usually watch high-definition television programs on large screens and smartphone users usually watch low resolution videos. Thus, we may conduct subjective studies on different devices. Based on the results of the studies, the users' typical video quality expectations x_u^* on each type of device can be deduced. Then, we can categorize video users according to their respective devices and provide differentiated QoE guarantees.

We define a finite set $\mathcal{G}=\{g_1,\dots,g_{|\mathcal{G}|}\}$ that represents different video quality expectations and assume $x_u^*\in\mathcal{G}, \forall u\in\cup_{t=1}^{\infty}\mathcal{U}^{\mathrm{av}}(t)$. Let $\mathcal{U}_j^{\mathrm{av}}(t)=\{u\in\mathcal{U}^{\mathrm{av}}(t):x_u^*=g_j\}$ denote the video users whose quality expectation is g_j. We consider the following constraints:

$$\mathrm{F}^{(2)}(g_j;\mathrm{q}_u)\le h_j, \forall g_j\in\mathcal{G}, \forall u\in\cup_{t=1}^{\infty}\mathcal{U}_j^{\mathrm{av}}(t), \tag{8}$$

where h_j is the QoE constraint for users with quality expectation g_j.

In sum, the goal of the admission control strategy and the rate adaptation algorithm is to maximize the number of users satisfying constraints (2)–(4), (7), or (8). In Section V, we introduce our rate adaptation algorithm and the admission control strategy when users' quality expectations are unknown. Then, in Section VI, we extend our rate adaptation and admission control algorithms to the case where each user's video quality expectation is known. For ease of reading, a summary of the notation used in this paper is given in Table II.

V. Rate Adaptation and Admission Control With Unknown Video Quality Expectation

If the quality expectations of video users are not known, we apply the same constraint on the second-order eCDFs of all video users. We first propose a rate adaptation algorithm and a corresponding admission control strategy. Then, we evaluate their performance via numerical simulation.

A. Rate Adaptation Algorithm

To clarify the design of our rate adaptation algorithm, we present an off-line problem formulation in which the future channel conditions and admission decisions are assumed to be known. Then, based on the analysis of this offline problem, we propose a new on-line rate adaptation algorithm.

If we consider a finite horizon T and assume that the realizations of channel conditions $\mathcal{C}(1), \ldots, \mathcal{C}(T)$ are known, the rate adaptation algorithm should solve the following feasibility problem:

$$\begin{aligned} &\text{find} && \mathbf{r}_{1:T} && \text{(9a)}\\ &\text{subject to:} && \mathbf{r}(t) \in \mathcal{C}(t) \cap \mathcal{R}(t) && \text{(9b)}\\ &&& \mathrm{q}_u(t) = \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t) && \text{(9c)}\\ &&& \mathrm{F}^{(2)}\left(x_i; \mathrm{q}_u\right) \le \mathrm{h}(x_i) &&\\ &&& \forall t \in \{1, \ldots, T\}, \forall x_i \in \mathcal{I}, \forall u \in \cup_{t=1}^{T} \mathcal{U}^{\mathrm{av}}(t). && \text{(9d)} \end{aligned}$$

The constraint (9b) is associated with the achievable rate region (2) and the available video source rates in (3). The constraint (9c) is because of the rate-quality model (4). The constraints (9d) are the QoE constraints (7) that were discussed in Section IV.D. For each admitted user, a series of QoE constraints are applied to the 2nd-order eCDF at discrete points in $\mathcal{I} = \left\{x_1, \cdots, x_{|\mathcal{I}|}\right\}$. Since the rate-quality function (9c) is concave, according to [37], the problem (9) is equivalent to the following convex optimization problem:

$$\begin{aligned} &\text{maximize}_{c\mathbf{r}_{1:T}, (\hat{\mathrm{q}}_u)_{1:T}} \; 0 && \text{(10a)}\\ &\quad\text{subject to:} \;\; \mathbf{r}(t) \in \mathcal{C}(t) \cap \mathcal{R}(t) && \text{(10b)}\\ &\qquad\qquad \hat{\mathrm{q}}_u(t) \le \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t) && \text{(10c)}\\ &\qquad\qquad \mathrm{F}^{(2)}\left(x_i; \hat{\mathrm{q}}_u\right) \le \mathrm{h}(x_i) &&\\ &\qquad\qquad \forall t \in \{1, \ldots, T\}, \forall x_i \in \mathcal{I}, \forall u \in \cup_{t=1}^{T} \mathcal{U}^{\mathrm{av}}(t). && \text{(10d)} \end{aligned}$$

where $\hat{\mathrm{q}}_u(t)$ are virtual variables introduced here to make the constraint (10c) convex. Note that the right-hand side of constraint (10c) equals $\mathrm{q}_u(t)$. For any $\hat{\mathrm{q}}_u(t)$ satisfying (10c), we have $\hat{\mathrm{q}}_u(t) \le \mathrm{q}_u(t)$. Since $\mathrm{F}^{(2)}(x; q)$ is decreasing in q, if $\hat{\mathrm{q}}_u(t)$ satisfies (10d), the constraint $\mathrm{F}^{(2)}\left(x_i; \mathrm{q}_u\right) \le \mathrm{h}(x_i)$ is satisfied as well.

By the definition in (5), the 2nd-order eCDF in constraint (10d) is determined by the entire process $\hat{\mathrm{q}}_u(\mathsf{A}_u), \ldots, \hat{\mathrm{q}}_u(\mathsf{D}_u)$. Due to the constraints (10b) and (10c), $\hat{\mathrm{q}}_u(t)$ depends on the rate $\mathrm{r}_u(t)$ and thus also depends on the rate region $\mathcal{C}(t)$. Therefore, the solution of (10) depends on the entire process $\mathcal{C}(1), \ldots, \mathcal{C}(T)$. In practice, the future channel conditions are unavailable to the rate-adaptation algorithm. In the following, we transform problem (10) to a simpler form that inspires our online rate adaptation algorithm.

Since (10) is a convex problem, if it is feasible, there exists a set of Lagrange multipliers $\lambda_{u,i}^* \ge 0$ for the constraints in (10d) such that a solution of (10) can be obtained by solving the following problem (see [38]):

$$\begin{aligned} &\underset{\mathbf{r}_{1:T}, (\hat{\mathrm{q}}_u)_{1:T}}{\text{maximize}} && \sum_{\forall u} \sum_{\forall x_i} \lambda_{u,i}^* \left[\mathrm{F}^{(2)}\left(x_i; \hat{\mathrm{q}}_u\right) - \mathrm{h}(x_i)\right] && \text{(11a)}\\ &\text{subject to:} && \mathbf{r}(t) \in \mathcal{C}(t) \cap \mathcal{R}(t) && \text{(11b)}\\ &&& \hat{\mathrm{q}}_u(t) \le \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t) &&\\ &&& \forall t \in \{1, \ldots, T\}, \forall u \in \cup_{t=1}^{T} \mathcal{U}^{\mathrm{av}}(t). && \text{(11c)} \end{aligned}$$

If we define a function $\mathrm{s}_{u,i}(t)$ as

$$\mathrm{s}_{u,i}(t) = \begin{cases} \frac{1}{\mathsf{T}_u}\left(\left[x_i - \hat{\mathrm{q}}_u(t)\right]^+ - \mathrm{h}(x_i)\right) & \text{if } \mathsf{A}_u \le t \le \mathsf{D}_u,\\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

the term $\mathrm{F}^{(2)}\left(x_i; \hat{\mathrm{q}}_u\right) - \mathrm{h}(x_i)$ in (11a) can be rewritten as

$$\begin{aligned} &\mathrm{F}^{(2)}\left(x_i; \hat{\mathrm{q}}_u\right) - \mathrm{h}(x_i)\\ =&\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \frac{1}{\mathsf{T}_u}\left(\left[x_i - \hat{\mathrm{q}}_u(t)\right]^+ - \mathrm{h}(x_i)\right)\\ =&\sum_{t=1}^{T} \mathrm{s}_{u,i}(t). \end{aligned} \tag{13}$$

Thus, $\mathrm{s}_{u,i}(t)$ indicates to what extent the variable $\hat{\mathrm{q}}_u(t)$ violates the constraint $\mathrm{F}^{(2)}\left(x_i; \hat{\mathrm{q}}_u\right) \le \mathrm{h}(x_i)$ in each slot. Substituting (13) in (11a) and changing the order of summation, the optimization in (11) becomes

$$\begin{aligned} &\underset{\mathbf{r}_{1:T}, (\hat{\mathrm{q}}_u)_{1:T}}{\text{maximize}} && \sum_{t=1}^{T}\left(\sum_{u \in \mathcal{U}^{\mathrm{av}}(t)} \sum_{x_i \in \mathcal{I}} \lambda_{u,i}^* \mathrm{s}_{u,i}(t)\right)\\ &\text{subject to:} && \mathbf{r}(t) \in \mathcal{C}(t) \cap \mathcal{R}(t)\\ &&& \hat{\mathrm{q}}_u(t) \le \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t)\\ &&& \forall t \in \{1, \ldots, T\}, \forall u \in \mathcal{U}^{\mathrm{av}}(t). \end{aligned} \tag{14}$$

Note that, except for the Lagrange multipliers, the optimization in (14) does not involve variables that depend on the entire process $\{\hat{\mathrm{q}}_u(t) : 1 \le t \le T\}$. Thus, (14) can be solved by minimizing the weighted sum $\sum_{u \in \mathcal{U}^{\mathrm{av}}(t)} \sum_{x_i \in \mathcal{I}} \lambda_{u,i}^* \mathrm{s}_{u,i}(t)$ in every slot. That is, if it is possible to estimate the Lagrange multiplier $\lambda_{u,i}^*$, then (10) can be solved by greedily choosing the rate vector $\mathbf{r}(t)$ at each time slot as the solution of the following problem:

$$\begin{aligned} &\underset{\mathbf{r}(t), \hat{\mathrm{q}}_u(t)}{\text{maximize}} && \sum_{u \in \mathcal{U}^{\mathrm{av}}(t)} \sum_{x_i \in \mathcal{I}} \lambda_{u,i}^* \mathrm{s}_{u,i}(t)\\ &\text{subject to:} && \mathbf{r}(t) \in \mathcal{C}(t) \cap \mathcal{R}(t)\\ &&& \hat{\mathrm{q}}_u(t) \le \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t)\\ &&& \forall u \in \mathcal{U}^{\mathrm{av}}(t). \end{aligned} \tag{15}$$

We introduce a method to approximate the Lagrange multiplier $\lambda_{u,i}^*$. We know that the Lagrange multiplier $\lambda_{u,i}^*$ indicates the difficulty in satisfying the constraint $\mathrm{F}_u^{(2)}(x_i;\hat{\mathrm{q}}_u) \le \mathrm{h}(x_i)$ [37]. Inspired by prior work in [39] and [40], we employ a virtual queue to capture this difficulty. For each admitted user u and each $x_i \in \mathcal{I}_u$, define the virtual queue as

$$\mathrm{v}_{u,i}(t) = \begin{cases} [\mathrm{v}_{u,i}(t-1)+\mathrm{s}_{u,i}(t)]^+ & \text{if } \mathsf{A}_u \le t \le \mathsf{D}_u, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

From (13) it follows that, if the summation of $\mathrm{s}_{u,i}(t)$ is large, then it is difficult to satisfy the constraint $\mathrm{F}_u^{(2)}(x_i;\hat{\mathrm{q}}_u) \le \mathrm{h}(x_i)$. The virtual queue captures the cumulative summation of $\mathrm{s}_{u,i}(t)$ up to slot t. Hence, the virtual queue reflects the level of difficulty in satisfying $\mathrm{F}_u^{(2)}(x_i;\hat{\mathrm{q}}_u) \le \mathrm{h}(x_i)$. Actually, for the special case where user set $\mathcal{U}^{\mathrm{av}}(t)$ is fixed for all t, it can be proved that the virtual queue asymptotically approaches $\lambda_{u,i}^*$ as $T \to \infty$ [39]. Hence, we replace the Lagrange multipliers in (15) with virtual queue $\mathrm{v}_{u,i}(t)$ and our online rate adaptation algorithm is summarized in Algorithm 1. In every slot, we maximize the weighted sum of $\mathrm{s}_{u,i}(t)$, where the weight is given by $\mathrm{v}_{u,i}(t-1)$. Thus users with larger virtual queues tend to be allocated more network resources. This helps users satisfy their QoE constraints.

Next, we introduce an admission control policy that is combined with our rate-adaptation algorithm to further improve performance.

Algorithm 1 Online algorithm for video data rate adaptation

```
for $t = 1 \to \infty$ do
    Choose rate vector $\mathbf{r}(t)$ that solves the problem
        $\underset{\mathbf{r}(t),\hat{\mathrm{q}}_u(t)}{\text{maximize}} \quad \sum_{u\in\mathcal{U}^{\mathrm{av}}(t)} \sum_{x_i\in\mathcal{I}} \mathrm{v}_{u,i}(t-1)\mathrm{s}_{u,i}(t)$
        subject to: $\mathbf{r}(t) \in \mathcal{C}(t) \cap \mathcal{R}(t)$
                    $\hat{\mathrm{q}}_u(t) \le \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t)$
                    $\forall u \in \mathcal{U}^{\mathrm{av}}(t),$                                   (17)
        where $\mathrm{s}_{u,i}(t) = \frac{1}{\mathsf{T}_u}\left([x_i - \hat{\mathrm{q}}_u(t)]^+ - \mathrm{h}(x_i)\right)$.
    $\forall u \in \mathcal{U}^{\mathrm{av}}(t), \forall x_i \in \mathcal{I}$, update virtual queues with
        $\mathrm{v}_{u,i}(t) = [\mathrm{v}_{u,i}(t-1) + \mathrm{s}_{u,i}(t)]^+$.
end for
```

B. Admission Control Strategy

Since a video stream typically has high data rate and thus consumes a large amount of resources, the arrival and departure of a single video user can have a significant impact on other video users' QoE. Our admission control strategy is designed to identify and block those video users who may consume excessive network resources. As has been discussed in Algorithm 1, resource allocation in each slot is determined by the solution of the optimization problem (17). Therefore, it is possible to estimate the QoE of a newly arrived user by solving (17) as if the user had already been admitted. Based on this idea, we propose a threshold-based admission control strategy, which is summarized in Algorithm 2. For each newly arrived video user $\bar{u}$, we first estimate its video quality $\bar{q}$ by solving the optimization problem (20), which is similar to the optimization problem (17). Then, we compare $\bar{q}$ with a threshold θ. If $\bar{q}$ is larger than θ, it is admitted to the network. Otherwise, it is rejected.

Algorithm 2 Admission control when video quality expectation is not known.

```
Inputs: Threshold $\theta$, admitted users $\mathcal{U}^{\mathrm{av}}(t-1)$, new user $\bar{u}$
1: Initialize video user set $\mathcal{U}^{\mathrm{av}+} \leftarrow \mathcal{U}^{\mathrm{av}}(t-1) \cup \{\bar{u}\}$
2: Estimate mean rate-quality parameters for all
   $u \in \mathcal{U}^{\mathrm{av}+}$:
       $\hat{\alpha}_u \leftarrow 1/\mathsf{T}_u \sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \alpha_u(t),$
       $\hat{\beta}_u \leftarrow 1/\mathsf{T}_u \sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \beta_u(t).$                    (18)
3: Initialize virtual queue for a new user $\bar{u}$:
   $\mathrm{v}_{\bar{u},i}(t-1) \leftarrow \frac{1}{|\mathcal{U}^{\mathrm{av}}(t-1)|} \sum_{u\in\mathcal{U}^{\mathrm{av}}(t-1)} \mathrm{v}_{u,i}(t-1), \forall i \in \mathcal{I}$
                                                                  (19)
4: Define variables
       $\hat{\mathbf{r}} = \left(\hat{r}_u : u \in \mathcal{U}^{\mathrm{av}+}\right)$
       $\hat{\mathbf{q}} = \left(\hat{q}_u : u \in \mathcal{U}^{\mathrm{av}+}\right)$
       $\hat{s}_{u,i} = \frac{1}{\mathsf{T}_u}\left([x_i - \hat{q}_u]^+ - \mathrm{h}(x_i)\right)$
   Find the solution $\mathbf{r}^* = (r_u^* : u \in \mathcal{U}^{\mathrm{av}+})$ of the
   optimization problem
       $\underset{\hat{\mathbf{r}},\hat{\mathbf{q}}}{\text{maximize}} \quad \sum_{u\in\mathcal{U}^{\mathrm{av}+}} \sum_{x_i\in\mathcal{I}} \mathrm{v}_{u,i}(t-1)\hat{s}_{u,i}$
       subject to: $\hat{\mathbf{r}} \in \mathcal{C} \cap \mathcal{R},$
                   $\hat{q}_u \le \hat{\alpha}_u \log(\hat{r}_u) + \hat{\beta}_u, \forall u \in \mathcal{U}^{\mathrm{av}+},$   (20)
   where the sets
       $\mathcal{C} = \{\mathbf{r} : \mathbb{E}[\mathrm{C}_t(\mathbf{r}) \le 0]\}$
   and
       $\mathcal{R} = \Pi_{u\in\mathcal{U}^{\mathrm{av}+}} \left[\frac{1}{\mathsf{T}_u}\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \mathrm{r}_u^{\min}(t), \frac{1}{\mathsf{T}_u}\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \mathrm{r}_u^{\max}(t)\right].$
5: Estimate the video quality delivered to new user via
       $\bar{q} = \hat{\alpha}_{\bar{u}} \log(r_{\bar{u}}^*) + \hat{\beta}_{\bar{u}}.$            (21)
6: If $\bar{q} > \theta$, admit the new user; otherwise, reject it.
```

The optimization problem (20) is different from (17) in the following three aspects. First, to predict the long-term QoE of users, we replace the instantaneous rate-quality parameters $\alpha_u(t)$ and $\beta_u(t)$ in (17) with the average rate-quality parameters $\hat{\alpha}_u$ and $\hat{\beta}_u$ (see the second step in Algorithm 2):

$$\hat{\alpha}_u = \frac{1}{\mathsf{T}_u}\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \alpha_u(t),$$

$$\hat{\beta}_u = \frac{1}{\mathsf{T}_u}\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \beta_u(t). \tag{22}$$

Second, we replace the instantaneous rate region $\mathcal{C}(t)$ with a rate region estimated by

$$\mathcal{C} = \{\mathbf{r} : \mathbb{E}[\mathrm{C}_t(\mathbf{r})] \leq 0\}. \tag{23}$$

Similarly, we replace the set of available video source rate $\mathcal{R}(t)$ by

$$\mathcal{R} = \Pi_{u\in\mathcal{U}^{\mathrm{av}+}} \left[\frac{1}{\mathsf{T}_u}\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \mathrm{r}_u^{\min}(t), \frac{1}{\mathsf{T}_u}\sum_{t=\mathsf{A}_u}^{\mathsf{D}_u} \mathrm{r}_u^{\max}(t)\right], \tag{24}$$

where $\mathcal{U}^{\mathrm{av}+} = \mathcal{U}^{\mathrm{av}}(t-1) \cup \{\bar{u}\}$. Third, for a newly arrived user $\bar{u}$, we initialize its virtual queue with the average virtual queues of the existing users (see the third step in Algorithm 2), i.e.,

$$\mathrm{v}_{\bar{u},i}(t-1) \leftarrow \frac{1}{|\mathcal{U}^{\mathrm{av}}(t-1)|} \sum_{u\in\mathcal{U}^{\mathrm{av}}(t-1)} \mathrm{v}_{u,i}(t-1), \forall i \in \mathcal{I}. \tag{25}$$

In (22), the rate-quality parameters $\alpha_u(\cdot)$ and $\beta_u(\cdot)$ are needed. For stored video streaming systems, the videos are pre-encoded. Thus, we assume the rate-quality parameters for the entire video stream are known. Also, the rate region $\mathcal{C}$ in (23) can be estimated using the time-average of the previous channel conditions. For example, in TDMA systems, we have $\mathrm{C}_t(\mathbf{r}) = \sum_{u\in\mathcal{U}^{\mathrm{av}}(t)} \frac{r_u}{\mathsf{P}_u(t)} + \sum_{u'\in\mathcal{U}^{\mathrm{p}}(t)} \frac{\mathsf{R}_{u'}(t)}{\mathsf{P}_{u'}(t)} - 1$ (see Section IV-B). Thus, we can estimate $\mathbb{E}\left[\frac{1}{\mathsf{P}_u}\right]$ and $\mathbb{E}\left[\sum_{u'\in\mathcal{U}^{\mathrm{p}}(t)} \frac{\mathsf{R}_{u'}(t)}{\mathsf{P}_{u'}(t)}\right]$ using the previous observations of the peak rate P_u and the high-priority users' data rate R_u. The estimated rate region is therefore $\mathcal{C} = \{\mathbf{r} : \sum_{u\in\mathcal{U}^{\mathrm{av}+}} \mathbb{E}\left[\frac{1}{\mathsf{P}_u}\right] r_u \leq 1 - \mathbb{E}\left[\sum_{u'\in\mathcal{U}^{\mathrm{p}}(t)} \frac{\mathsf{R}_{u'}(t)}{\mathsf{P}_{u'}(t)}\right]\}$.

As was discussed in Section V-A, the virtual queue $\mathrm{v}_{u,i}(t)$ captures the difficulty for an admitted video user to satisfy the QoE constraints. Thus, users with large virtual queues tend to be allocated more network resources. In other words, the virtual queues drive the priorities in resource allocation. Because it is difficult to estimate the length of a virtual queue before a new user is admitted, we simply initialize the virtual queue of the newly arrived user with the average virtual queues of all existing video users. In this way, we actually estimate the video quality $\bar{q}$ when an average priority is assigned to the new user. Next, we introduce an approach to optimize the admission threshold θ in Algorithm 2.

C. Online Algorithm for Threshold Optimization

Denote by $\mathrm{g}(\theta)$ the probability that a video user's QoE constraints are satisfied when the threshold is θ. Also, denote by $\mathrm{e}(\theta)$ the probability that a video user is admitted into the network but its QoE constraints are not satisfied. Our goal is to find the threshold θ^* that maximizes $\mathrm{g}(\theta)$. We have conducted extensive simulations under different channel conditions and QoE constraints. From all the simulation results, we observed that the optimal threshold θ^* maximizes $\mathrm{g}(\theta)$ if

$$\begin{cases} \mathrm{e}(\theta) > 0, & \forall \theta < \theta^* \\ \mathrm{e}(\theta) = 0, & \forall \theta \geq \theta^* \end{cases} \tag{26}$$

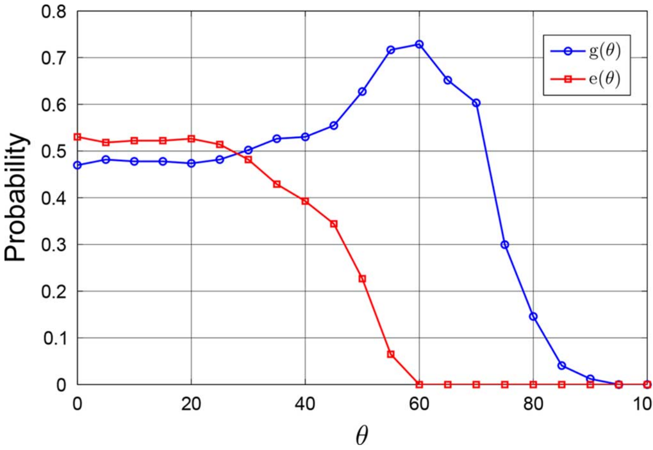


Fig. 5. The two plots show (i) the percentage of video users who satisfy the QoE constraints and (ii) the percentage of video users who are admitted into the network but do not satisfy the QoE constraints under different admission control thresholds.

This means that $\mathrm{g}(\theta)$ is maximized if θ is just large enough to make all the admitted users satisfy the QoE constraints. As an example, using the same simulation configurations that are detailed later in Section V-D, we simulated and plotted the functions $\mathrm{e}(\theta)$ and $\mathrm{g}(\theta)$ in Fig. 5. It is seen that $\theta^* = 60$ satisfies (26) and $\mathrm{g}(\theta)$ is also maximized at θ^*. Therefore, to find the optimal threshold θ^*, it is sufficient to find a threshold that satisfies (26).

We propose an iterative algorithm that automatically adjusts the threshold to θ^*. This is summarized in Algorithm 3. In each iteration, the algorithm observes the 2^{nd}-order eCDFs of L video users who have been admitted into the network since the end of the last iteration. Then the algorithm updates the threshold via

$$\theta^{n+1} = \theta^n + \epsilon^n y^n, \tag{27}$$

where θ^n denote the admission control threshold in the n^{th} iteration. The value $y^n \in \{-1, 1\}$ determines whether to increase or to decrease the threshold. The quantity $\epsilon^n > 0$ is an updating step size. If the algorithm observes a video user whose 2^{nd}-order eCDF violates the QoE constraints, then it is probable that $\mathrm{e}(\theta^n) > 0$ and $\theta^n < \theta^*$. Therefore, the algorithm increases the threshold by setting $y^n = 1$. Otherwise, if all the L video users satisfy the constraints, the threshold is possibly larger than θ^*. Thus the algorithm decreases the threshold by setting $y^n = -1$. The updating step size is

$$\epsilon^n = \epsilon^0/m, \tag{28}$$

where m counts the number of sign changes in the series $\{y^1, \ldots, y^n\}$ (see step 8) and ϵ^0 is the initial step-size. Here, m is introduced to accelerate the convergence of the algorithm. The reason is as follows: If θ^n is far from θ^*, the sign of y^n does not change frequently and m increases slowly. Thus, the step-size ϵ^n stays large and θ^n moves towards θ^* quickly. When θ^n is moved to a small neighborhood of θ^*, the sign of y^n changes frequently and thus m increases rapidly. Therefore, the step-size ϵ^n decreases to zero quickly, which makes θ^n converge.

Algorithm 3 The threshold optimization algorithm when the video quality expectation is unknown.

```
Inputs: $L = 100$, initial threshold $\theta^0 = 0$, initial step-size $\epsilon^0 = 10$, and initial counter $m = 1$
for $n = 1 \to \infty$ do
    Observe the $2^{\mathrm{nd}}$-order eCDFs of $L$ admitted video users.
    if there exits a user that does not satisfy the QoE constraints then
        $y^n \leftarrow 1$
    else
        $y^n \leftarrow -1$
    end If
    If $y^n \neq y^{n-1}$, $m \leftarrow m + 1$.
    Update threshold with
        $\theta^{n+1} = \theta^n + \epsilon^n y^n$,    (29)
    where $\epsilon^n = \epsilon^0 / m$.
end for
```

In the following, we analyze the convergence of Algorithm 3. Based on our observations from the simulations, we make the following assumption:

Assumption 1: The function $\mathrm{e}(\theta)$ is a continuous function and is strictly decreasing on $[0, \theta^*]$.

We define $\mathrm{p}^L(\theta)$ as the probability that all the L admitted video users in an iteration satisfy the QoE constraints. Since increasing the threshold θ would block more users and thus reserve more network resources to the admitted users, we assume that $\mathrm{p}^L(\theta)$ is a continuously increasing function of θ. Also, according to Assumption 1, when $\theta > \theta^*$, all the admitted users satisfy the QoE constraints and thus $\mathrm{p}^L(\theta) = 1$. Thus, we have the following assumption on $\mathrm{p}^L(\theta)$:

Assumption 2: The function $\mathrm{p}^L(\theta)$ is a continuous and increasing function of θ. For all $\theta > \theta^*$, we have $\mathrm{p}^L(\theta) = 1$. Furthermore, assume that there exists a constant $M > 0$ such that $|\mathrm{p}^L(\theta) - \mathrm{p}^L(\theta')| \geq M|\theta - \theta'|$ for all $\theta' < \theta$ and $\mathrm{p}^L(\theta) < 1$.

The following theorem assures that if L is sufficiently large, θ^n converges to an arbitrarily small neighborhood of θ^* as $n \to \infty$. Its proof is given in Appendix B.

Theorem 2: Let $\delta > 0$ be an arbitrarily small number. If Assumptions 1 and 2 are satisfied and $L \geq \frac{-\log 2}{\log(1 - \mathrm{e}(\theta^* - \delta))}$, then θ^n converges as $n \to \infty$ and $\lim_{n \to \infty} \theta^n \in [\theta^* - \delta, \theta^*]$.

D. Simulation Results

Below, we evaluate our rate-adaptation algorithm and the admission control strategy via numerical simulations. We assume the duration of a time slot is $\Delta T = 1$ second. The high-priority users' arrivals follow a Poisson process with average arrival rate of $\frac{1}{20}$ users/second. The time spent by a high-priority user in the network is exponentially distributed with a mean value of 200 seconds. Video users also arrival as a Poisson process with a average arrival rate of $\frac{1}{20}$ users/second. Since video streams are typically more than tens of seconds long, we assume the time spent by a video user in the network is at least 40 seconds. In particular, for all video users, we set $\mathsf{T}_u = \max\{\mathsf{T'}_\mathrm{u}, 40\}$, where $\mathsf{T'}_u$ is exponentially distributed with a mean value of 200 seconds.

To simulate variations of the rate-quality characteristics in each video stream, we assume the rate-quality parameters $(\alpha_u(t), \beta_u(t))$ of each slot are independently sampled from the rate-quality parameters in the video database [1]. We assume the minimum and maximum available data rate for video users in (3) are $\mathrm{r}_u^{\min}(t) = 302$ kbps and $\mathrm{r}_u^{\max}(t) = 6412$ kbps, which are the minimum and maximum rate of the videos in the database [1]. For high-priority users, the downloading data rate R_u is assumed to be uniformly distributed in $[100, 300]$ kbps.

We assume the wireless system is a TDMA system. The rate region $\mathcal{C}(t)$ is that introduced in Section IV.B. We model the peak transmission rate $\mathsf{P}_u(t)$ as the product of two independent random variables, i.e., $\mathsf{P}_u(t) = \mathsf{P}_u^{\mathrm{avg}} \times \mathsf{P}_u^*(t)$. The random variable $\mathsf{P}_u^{\mathrm{avg}}$ is employed to simulate the heterogeneity of channel condition across users and remains constant during a user's sojourn. We assume that $\mathsf{P}_u^{\mathrm{avg}}$ is uniformly distributed on $[1250\gamma, 3750\gamma]$ kbps, where the parameter γ is used to scale the channel capacity in our simulations. The random variable $\mathsf{P}_u^*(t)$ is employed to simulate channel variation across time slots. We assume that $\{\mathsf{P}_u^*(t) : t \in \mathbb{N}^+\}$ is an i.i.d. process with $\mathsf{P}_u^*(t)$ being uniformly distributed on $[0.5, 1.5]$. In our simulations, we set $\mathcal{I} = \{30, 40, 50, 60, 70\}$. Correspondingly, for $x_i = 30, 40, 50, 60$, and 70, we let the constraints $\mathrm{h}(x_i) = 0.7, 1.0, 3.0, 7.0$ and 15.0, respectively.

We first evaluate the performance of the rate-adaptation algorithm when admission control is not applied. We set the scaling parameter $\gamma = 12$ and simulate Algorithm 1 until 100 users have arrived and departed the network. We plot the 2^{nd}-order eCDFs of the video users in Fig. 6(a). It may be seen that, using Algorithm 1, the 2^{nd}-order eCDFs of the video users all satisfy the constraints. By comparison, if we adapt the rate vector to maximize the sum of the average-quality of all users[1], the QoE constraint is violated by many users (see Fig. 6(b)).

Next, we evaluate the proposed admission control strategy. In Fig. 7(a), we fix the channel scaling parameter to be $\gamma = 6$ and plot the threshold θ^n at every iteration of the online threshold optimization algorithm. Recall that the optimal threshold is $\theta^* = 60$ (see Fig. 5). Fig. 7(a) shows that θ^n converges to θ^* after 200 video user arrivals. We have assumed that the average arrival rate of video users is $1/20$ users/seconds. Thus, 200 video user arrivals require about 200×20 seconds $= 1.1$ hours. Since our goal is to optimize the performance of the network in the long run, this convergence speed is acceptable.

We scale the channel scaling parameter from $\gamma = 6$ to $\gamma = 16$. The percentage of video users whose video qualities satisfy the QoE constraints is shown in Fig. 7(b). When compared with the average-quality-maximized rate-adaptation algorithm, the percentage of video users who satisfy the QoE constraints is improved significantly even if no admission control is applied. The admission control policy further improves the performance especially when the channel condition is poor. For example, at $\gamma = 6$, the proposed algorithms satisfy the QoE constraints of 70% of the video users while the average-quality-maximized rate-adaptation algorithm only satisfies the constraints of 20% of the video users. At a moderate channel condition of $\gamma = 12$, about 77% of the video users satisfy the QoE constraints when the average-quality maximizing algorithm is applied. The proposed algorithms achieve the same performance at $\gamma = 7.5$, reducing the consumption of resources by $(12 - 7.5)/12 = 38\%$.

[1] This is achieved by maximizing $\sum_{u \in \mathcal{U}^{\mathrm{av}}(t)} \mathrm{q}_u(t)/\mathsf{T}_u$ in each slot.

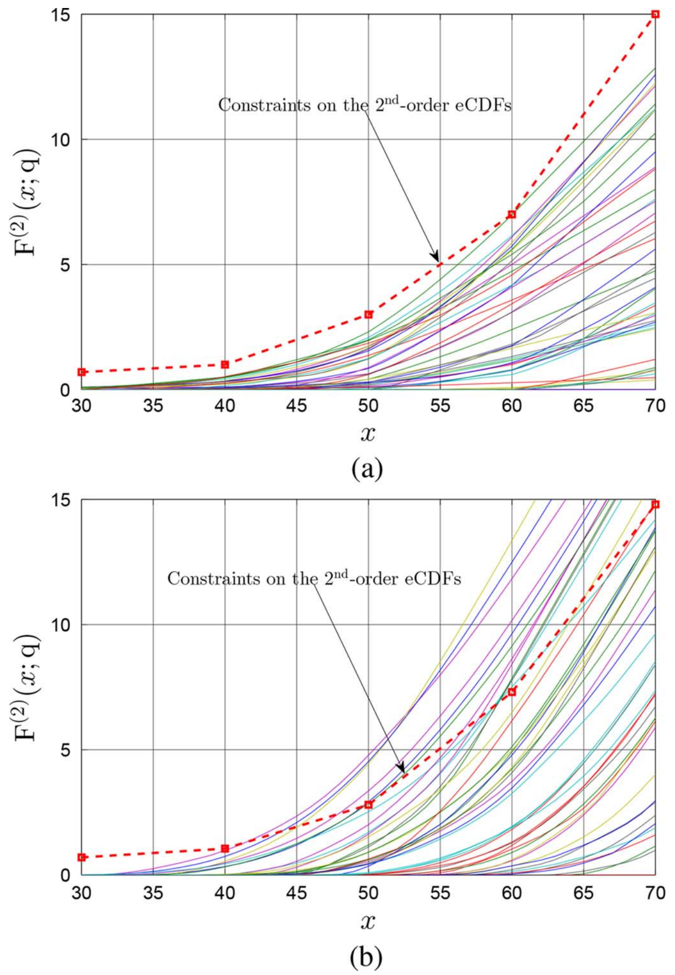


Fig. 6. Simulation results of rate-adaptation algorithms when admission control is not applied. (a) The 2^{nd}-order eCDFs of the video users when the proposed rate-adaptation is used. (b) The 2^{nd}-order eCDFs of the video users when the rate vector is adapted to maximize the sum of users' video qualities. (a) Proposed rate-adaptation algorithm, (b) Average-quality maximized rate adaptation.

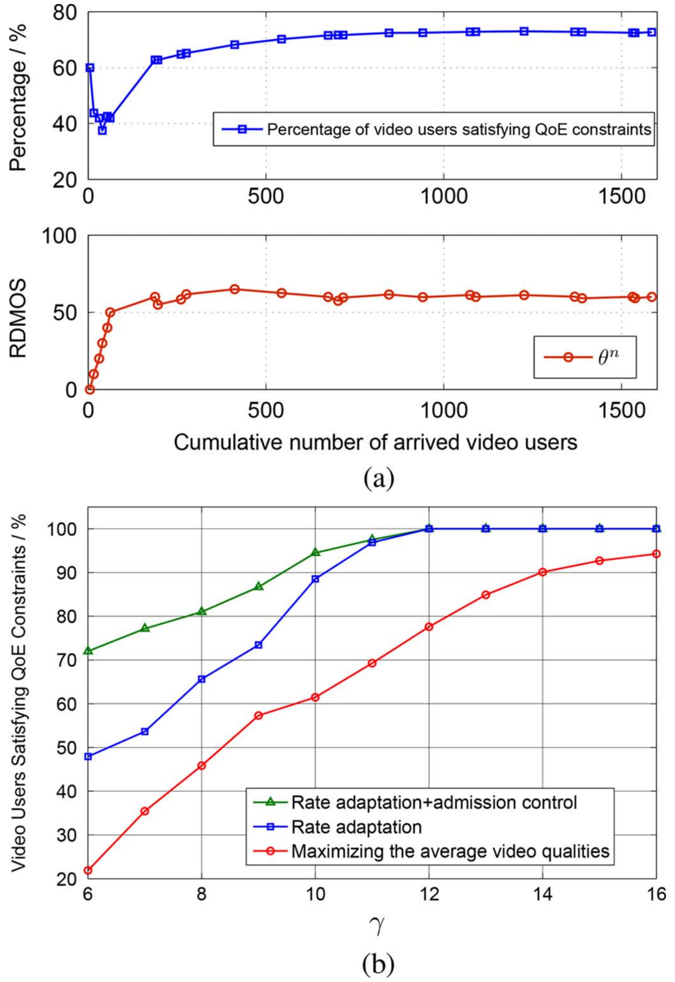


Fig. 7. (a) The performance of the proposed admission control strategy when the scaling parameter is $\gamma = 6$. (b) Simulation results of the proposed algorithms under different channel scaling parameters. Each data point on the figure is obtained by simulating 2000 video user arrivals.

VI. Rate Adaptation and Admission Control With Known Video Quality Expectation

In this section, we extend the rate adaptation algorithm and the admission control strategy to the case where the video quality expectation of each user is known. We first explain the extended algorithms and then evaluate their performance via simulation.

A. The Extended Rate Adaptation and Admission Control Algorithms

In Section IV-D, we defined the finite set $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ to represent different video quality expectations among video users. In the following, we call users with $x_u^* = g_j \in \mathcal{G}$ the Type-j users. Each Type-j video user needs only satisfy one QoE constraint, i.e.,

$$\mathrm{F}^{(2)}(g_j;\mathrm{q}_u) \le h_j. \tag{30}$$

Thus, we extend the rate adaptation method in Algorithm 1 by maintaining one virtual queue for each user. In particular, the virtual queue of a Type-j user u is defined as

$$\mathrm{v}_u(t) = \begin{cases} [\mathrm{v}_u(t-1) + \mathrm{s}_u(t)]^+ & \text{if } \mathsf{A}_u \le t \le \mathsf{D}_u, \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

where

$$\mathrm{s}_u(t) = \begin{cases} \frac{1}{\mathsf{T}_u}\left([g_j - \hat{\mathrm{q}}_u(t)]^+ - h_j\right) & \text{if } \mathsf{A}_u \le t \le \mathsf{D}_u, \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

In each slot, the rate vector $\mathbf{r}(t)$ is adapted by solving

$$\begin{aligned} &\underset{\mathbf{r}(t),\hat{\mathrm{q}}_u(t)}{\text{maximize}} && \sum_{u\in\mathcal{U}^{\mathrm{av}}(t)} \mathrm{v}_u(t-1)\mathrm{s}_u(t) \\ &\text{subject to:} && \mathbf{r}(t) \in \mathcal{C}(t) \cap \mathcal{R}(t), \\ &&& \hat{\mathrm{q}}_u(t) \le \alpha_u(t)\log(\mathrm{r}_u(t)) + \beta_u(t) \\ &&& \forall u \in \mathcal{U}^{\mathrm{av}}(t). \end{aligned} \tag{33}$$

For admission control, we extend Algorithm 2 by applying different thresholds to different types of users. In particular, for a newly arrived Type-j video user $\bar{u}$, we initialize its virtual queue by averaging the virtual queues of all existing video users i.e.,

$$\mathrm{v}_{\bar{u}}(t-1) \leftarrow \frac{1}{|\mathcal{U}^{\mathrm{av}}(t-1)|} \sum_{u\in\mathcal{U}^{\mathrm{av}}(t-1)} \mathrm{v}_u(t-1). \qquad (34)$$

Letting $\mathcal{U}^{\mathrm{av}+} = \mathcal{U}^{\mathrm{av}}(t) \cup \{\bar{u}\}$, we define variables $\hat{\mathbf{r}} = \left(\hat{r}_u : u \in \mathcal{U}^{\mathrm{av}+}\right)$, $\hat{\mathbf{q}} = \left(\hat{q}_u : u \in \mathcal{U}^{\mathrm{av}+}\right)$, and $\hat{s}_u = \frac{1}{\mathsf{T}_u}\left([g_j - \hat{q}_u]^+ - h_j\right)$. We then find the solution $\mathbf{r}^* = \left(r_u^* : u \in \mathcal{U}^{\mathrm{av}+}\right)$ of the following problem:

$$\begin{aligned} \underset{\hat{\mathbf{r}},\hat{\mathbf{q}}}{\text{maximize}} \quad & \sum_{u\in\mathcal{U}^{\mathrm{av}+}} \mathrm{v}_u(t-1)\hat{s}_u \\ \text{subject to:} \quad & \hat{\mathbf{r}} \in \mathcal{C} \cap \mathcal{R}, \\ & \hat{q}_u \le \hat{\alpha}_u \log(\hat{r}_u) + \hat{\beta}_u \\ & \forall u \in \mathcal{U}^{\mathrm{av}+} \end{aligned} \qquad (35)$$

where $\hat{\alpha}_u$, $\hat{\beta}_u$, $\mathcal{C}$, and $\mathcal{R}$ are given by (22), (23), and (24), respectively. Finally, we estimate the video quality of the new user by $\bar{q} = \hat{\alpha}_{\bar{u}} \log(r_{\bar{u}}^*) + \hat{\beta}_{\bar{u}}$ and compare $\bar{q}$ with a threshold θ_j to make the admission decision.

Next, we discuss how to optimize the threshold θ_j for Type-j users.

B. The Extended Threshold Optimization Algorithm

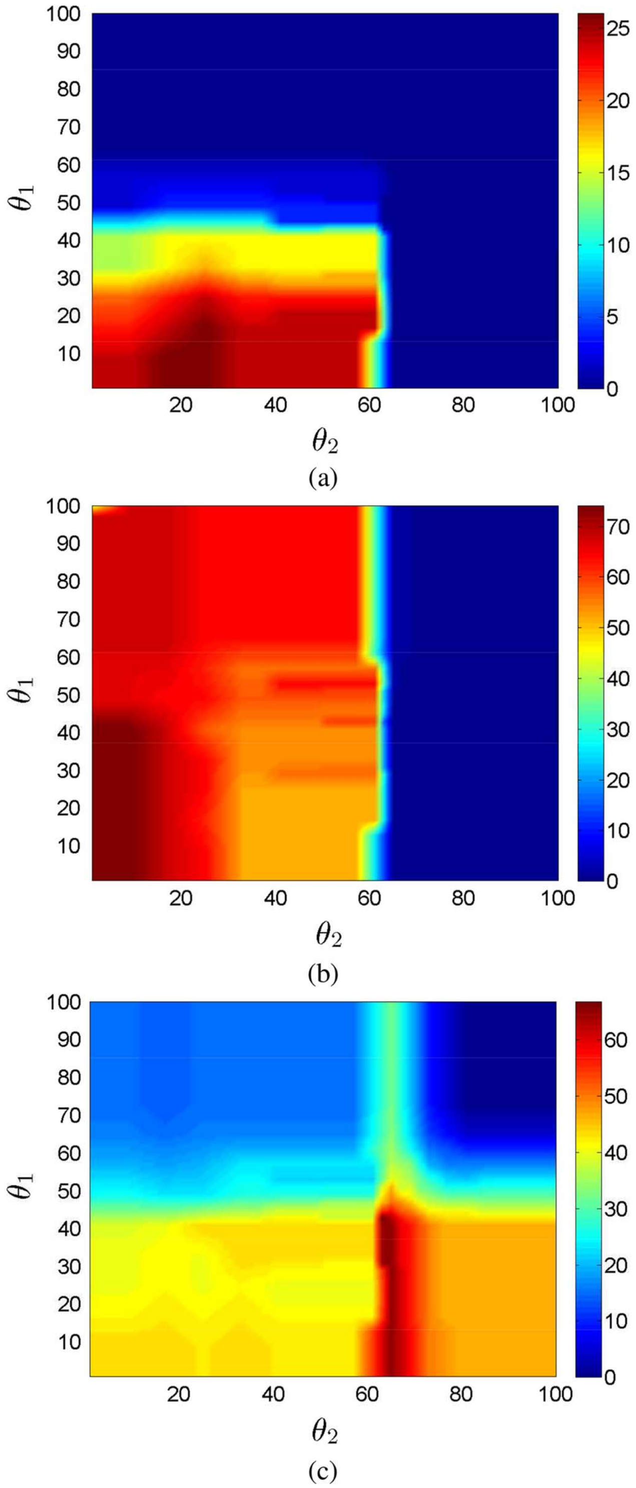
Denote by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{|\mathcal{G}|})$ the vector of thresholds for all types of video users. Define $\mathrm{g}(\boldsymbol{\theta})$ to be the probability that a video user satisfies the QoE constraints when the threshold vector is $\boldsymbol{\theta}$. Also, define $\mathrm{e}_j(\boldsymbol{\theta})$ as the probability that a Type-j video user's QoE constraint is not satisfied. To determine the optimal threshold vector $\boldsymbol{\theta}$ that maximizes $\mathrm{g}(\boldsymbol{\theta})$, we ran simulations under a variety of relevant channel conditions and QoE constraints. We found that a threshold vector $\boldsymbol{\theta}^* = \left(\theta_1^*, \ldots, \theta_{|\mathcal{G}|}^*\right)$ maximizes $\mathrm{g}(\boldsymbol{\theta})$ if

$$\begin{cases} \mathrm{e}_j(\boldsymbol{\theta}) > 0, & \forall \boldsymbol{\theta} \prec \boldsymbol{\theta}^* \\ \mathrm{e}_j(\boldsymbol{\theta}) = 0, & \forall \boldsymbol{\theta} \succeq \boldsymbol{\theta}^*. \end{cases}, \forall 1 \le j \le |\mathcal{G}|. \qquad (36)$$

Here, the partial order $\boldsymbol{\theta} \prec \boldsymbol{\theta}^*$ indicates that $\boldsymbol{\theta} \neq \boldsymbol{\theta}^*$ and $\theta_j \le \theta_j^*, \forall j$. The partial order $\boldsymbol{\theta} \succeq \boldsymbol{\theta}^*$ indicates that $\theta_j \ge \theta_j^*, \forall j$. The condition in (36) means that if $\boldsymbol{\theta}^*$ is an optimal threshold vector and we increase all entries of $\boldsymbol{\theta}^*$, the QoE constraints of all the admitted users can still be satisfied. Conversely, if we decrease all the entries of $\boldsymbol{\theta}^*$, the QoE constraints of all types of users will be violated with a non-zero probability. To illustrate this, we considered two types of video users who arrive to the network with equal probability and simulated the function $\mathrm{e}_1(\boldsymbol{\theta})$, $\mathrm{e}_2(\boldsymbol{\theta})$, and $\mathrm{g}_1(\boldsymbol{\theta})$ using the same setting as in Section V-D. From Figs. 8(a) and 8(b), it can be seen that the $\boldsymbol{\theta}$s in $[0, 42] \times [63, 64]$ satisfy the condition (36) because they lie on the boundary of the region where $\{\mathrm{e}_1(\boldsymbol{\theta}) > 0 \text{ and } \mathrm{e}_2(\boldsymbol{\theta}) > 0\}$. From Fig. 8(c), it is seen that the function $\mathrm{g}(\boldsymbol{\theta})$ is also maximized on the region $[0, 42] \times [63, 64]$.

We devise an iterative algorithm to find the threshold vector $\boldsymbol{\theta}^*$ satisfying (36). Denote by $\boldsymbol{\theta}^n$ the threshold vector at the n^{th} iteration, the algorithm observes the 2^{nd}-order eCDFs of L admitted video users and updates the threshold vector using

$$\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \mathrm{diag}(\boldsymbol{\epsilon}^n)\mathbf{y}^n, \qquad (37)$$



(a)

(b)

(c)

Fig. 8. (a) The probability of admitted Type-1 video users whose QoE constraints are violated. (b) The probability of admitted Type-2 video users whose QoE constraints are violated. (c) The probability of video users whose QoE constraints are satisfied. All results are shown in percentages. We assume these two types of user have $x_u^* = 40$ and 60, respectively. The QoE constraint in (30) is assumed to be $h_1 = h_2 = 1$. The channel scaling parameter is set as $\gamma = 6$. (a) $\mathrm{e}_1(\boldsymbol{\theta})$, (b) $\mathrm{e}_2(\boldsymbol{\theta})$, (c) $\mathrm{g}(\boldsymbol{\theta})$.

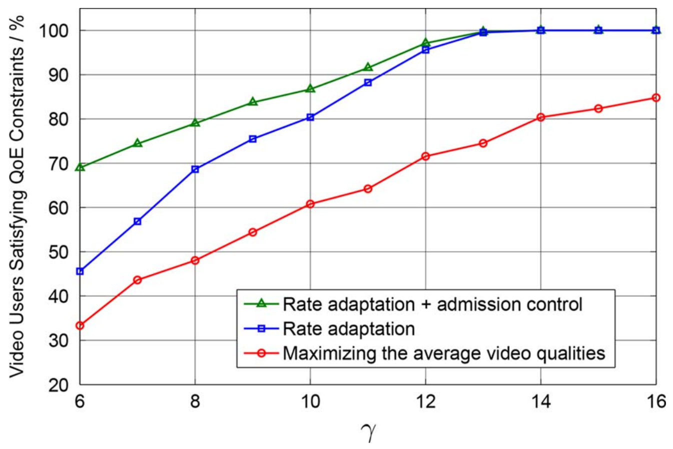


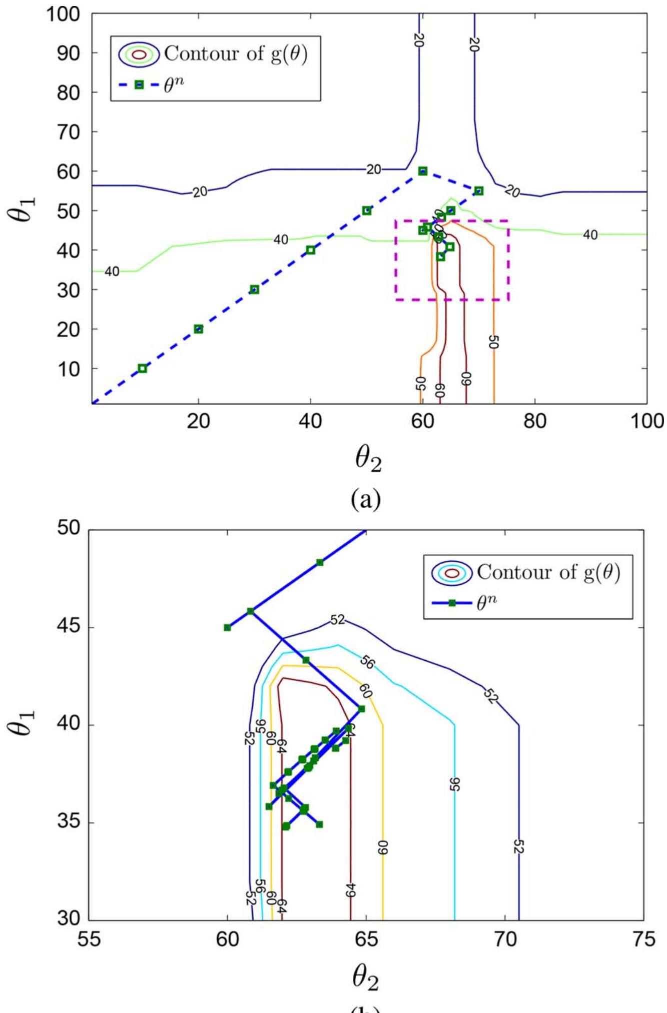
Fig. 9. Simulation results of the proposed admission control policy under different channel scaling parameters. Three cases are simulated. (1): rate adaptation is applied to maximize the average video quality and all users are admitted; (2): the proposed rate adaptation are applied and all users are admitted; (3): both the proposed rate adaptation and admission control algorithm are applied.

where $\mathrm{diag}(\boldsymbol{\epsilon}^n)$ is the diagonal matrix with diagonal entries being $\epsilon_1, \ldots, \epsilon_{|\mathcal{G}|}$. In (37), $\mathbf{y}^n = \left(y_1^n, \ldots, y_{|\mathcal{G}|}^n\right)$ is a $|\mathcal{G}|$-dimensional vector where $y_j^n \in \{-1, 1\}$ is the updating direction for θ_j. The vector $\boldsymbol{\epsilon} = \left(\epsilon_1^n, \ldots, \epsilon_{|\mathcal{G}|}^n\right)$ is the corresponding update step-size. Among the L video users, if a Type-j video user's QoE constraint is not satisfied, the algorithm sets $y_j^n = 1$. Otherwise, if all the Type-j video users' QoE constraints are satisfied, the algorithm sets $y_j^n = -1$. The step-size ϵ_j^n is given by $\epsilon_j^n = \epsilon_j^0 / m_j$, where m_j counts the sign changes in $\left\{y_j^1, \ldots, y_j^n\right\}$ and ϵ_j^0 is the initial step-size. Next, we show the performance of our rate adaptation algorithm and the admission control strategy via simulation.

C. Simulation Results

In our simulations, we assume that there are two types of video users. Both types of video users arrive as a Poisson process with arrival rate $1/40$ users/second. We assume that Type-1 users have $x_u^* = 40$ while Type-2 users have $x_u^* = 60$. We also set $h_1 = h_2 = 1$. In Fig. 9, we plot the percentage of video users whose video qualities satisfy the QoE constraint (8). It can be seen that, for all tested channel scaling parameters, our rate adaptation algorithm outperforms the average-quality maximizing algorithm. The percentage of video users who satisfy the QoE constraints improved significantly even when the admission control strategy was not applied. At $\gamma = 12$, about 71% of video users satisfied the QoE constraints when the average-quality maximizing algorithm was applied. The proposed algorithms achieve the same performance at $\gamma = 6.5$. Thus, the proposed algorithms reduce the consumption of resources by $(12 - 6.5)/12 = 46\%$.

In Fig. 10, we plot the threshold vectors $\boldsymbol{\theta}^n$ in the proposed threshold optimizing algorithm when the channel scaling parameter is fixed at $\gamma = 6$. We set the initial updating step-size $\epsilon_j^0 = 10, \forall j$. It is apparent that the threshold vector converges quickly to the area where $\mathrm{g}(\boldsymbol{\theta})$ is maximized.



(a)

(b)

Fig. 10. The updated threshold vector $\boldsymbol{\theta}^n$s of the proposed threshold optimization algorithm are shown in (a). The dashed box is shown blown up in (b) to illustrate more detail. The contours of $\mathrm{g}(\boldsymbol{\theta})$ are also shown on the figure for reference. (a) Convergence performance of the admission control strategy, (b) A zoom-in view.

VII. Conclusions and Future Work

We created a new QoE metric based on the second-order empirical cumulative distribution function (eCDF) of time-varying video quality. We then proposed an online rate adaptation algorithm to maximize the percentage of video users who satisfy the QoE constraints on the second-order cumulative distribution function. Furthermore, we devised a threshold-based admission control strategy that blocks new video users whose QoE constraints cannot be satisfied. Simulation results showed that combining the proposed approaches leads to a 40% reduction in wireless network resource consumption.

The users' video quality expectation play a critical role in our QoE metric. It is important to observe that our subjective study was conducted in a controlled environment. In reality, however, users' expectations for video quality may depend on various factors in the environment (e.g., user mobility, device type, lighting conditions). In the future, we plan to conduct subjective study in more diversified, "worldly" environments to obtain a better understanding of and an improved ability to predict users' quality expectations.

Appendix A
Proof of Theorem 1

Proof: Since $[x - \mathrm{q}(t)]^+$ is a convex function of x, $\mathrm{F}^{(2)}(x;\mathrm{q})$ is a linear combination of $[x - \mathrm{q}(t)]^+$ and is thus also a convex function of x. Without loss of generality, assume function $\mathrm{h}(x)$ is also convex[2]. Let $x_i < x_j$, where $i, j \in \mathcal{I}$. If (7) is satisfied, then $\mathrm{F}^{(2)}(x_i;\mathrm{q}) \leq \mathrm{h}(x_i)$ and $\mathrm{F}^{(2)}(x_j;\mathrm{q}) \leq \mathrm{h}(x_j)$. For any $\lambda \in [0,1]$ and $x = \lambda x_i + (1-\lambda)x_j$, we have $\mathrm{F}^{(2)}(x;\mathrm{q}) = \mathrm{F}^{(2)}(\lambda x_i + (1-\lambda)x_j;\mathrm{q}) \leq \lambda \mathrm{F}^{(2)}(x_i;\mathrm{q}) + (1-\lambda)\mathrm{F}^{(2)}(x_j;\mathrm{q}) \leq \lambda \mathrm{h}(x_i) + (1-\lambda)\mathrm{h}(x_j) = \mathrm{h}(x)$. Because $[0,100]$ is a compact set, the convexity of $\mathrm{h}(x)$ implies its continuity. Therefore, $\mathrm{h}(x)$ can be approximated by piece-wise linear functions to arbitrary accuracy. ■

Appendix B
Proof of Theorem 2

Proof: Note that Algorithm 3 can be viewed as a stochastic approximation algorithm [41] with an associated mean ordinary differential equation (ODE)

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = \mathbb{E}[\mathsf{Y}(\theta(t))], \tag{38}$$

where $\mathsf{Y}(\theta)$ is a random variable that denotes the updating direction when the threshold is θ, we have

$$\mathbb{E}[\mathsf{Y}(\theta(t))] = 1 - 2\mathrm{p}^L(\theta(t)). \tag{39}$$

According to Assumption 2, there exists a unique θ' such that $\mathrm{p}^L(\theta') = 1/2$ and $\mathbb{E}[\mathsf{Y}(\theta')] = 0$. By the monotonicity of $\mathrm{p}^L(\theta)$, we have $\mathbb{E}[\mathsf{Y}(\theta)] > 0, \forall \theta < \theta'$ and $\mathbb{E}[\mathsf{Y}(\theta)] < 0, \forall \theta > \theta'$. If we define a function $\mathrm{V}(\theta) = 1/2(\theta - \theta')^2$, then

$$\begin{aligned}\frac{\mathrm{dV}(\theta(t))}{\mathrm{d}t} &= (\theta(t) - \theta')\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} \\ &= (\theta(t) - \theta')\left(1 - 2\mathrm{p}^L(\theta(t))\right) \\ &= -(\theta(t) - \theta')\left(2\mathrm{p}^L(\theta(t)) - 2\mathrm{p}^L(\theta')\right) \\ &= -2(\theta(t) - \theta')\left(\mathrm{p}^L(\theta(t)) - \mathrm{p}^L(\theta')\right)\end{aligned}$$

For all $\theta < \theta'$, we have $\mathrm{p}^L(\theta) - \mathrm{p}^L(\theta') \leq M(\theta - \theta')$. For all $\theta > \theta'$, $\mathrm{p}^L(\theta) - \mathrm{p}^L(\theta') \geq M(\theta - \theta')$. In sum, we have

$$\frac{\mathrm{dV}(\theta(t))}{\mathrm{d}t} \leq -2M(\theta(t) - \theta')^2.$$

By Theorem 5.4.1 in ([41], p.145), we have $\lim_{n\to\infty} \theta^n = \theta'$. Next, we prove that $\theta' \in [\theta^* - \delta, \theta^*)$.

We define a binary random variable S_u such that $\mathsf{S}_u = 1$ if video user u satisfies the QoE constraints $\{\mathrm{F}^{(2)}(x_i;\mathrm{q}_u) \leq h_i, \forall x_i \in \mathcal{I}\}$. Otherwise, we define $\mathsf{S}_u = 0$. Denote by $\mathcal{U}^L = \{u_1, \ldots, u_L\}$ the indices of the L admitted video users in an iteration of Algorithm 2. Let π_θ be the joint distribution of the variables $\{\mathsf{S}_u, \forall u \in \mathcal{U}^L\}$ when the admission threshold is θ. Then, we have

$$\begin{aligned}\mathrm{p}^L(\theta) &= \mathbb{P}^{\pi_\theta}\left(\mathsf{S}_u = 1, \forall u \in \mathcal{U}^L\right) \\ &= \mathbb{P}^{\pi_\theta}\left(\mathsf{S}_{u_1} = 1\right)\mathbb{P}^{\pi_\theta}\left(\mathsf{S}_{u_\ell} = 1, 2 \leq \ell \leq L | \mathsf{S}_{u_1} = 1\right).\end{aligned} \tag{40}$$

Since the users are competing with each other for network resources, if the QoE constraints of a video user are satisfied, the probability of satisfying other users' QoE constraints is reduced. Thus,

$$\begin{aligned}&\mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_\ell} = 1, \forall 2 \leq \ell \leq L | \mathsf{S}_1 = 1) \\ \leq &\mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_\ell} = 1, \forall 2 \leq \ell \leq L).\end{aligned} \tag{41}$$

Substitute (41) into (40) yields

$$\begin{aligned}\mathrm{p}^L(\theta) &\leq \mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_1} = 1)\mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_\ell} = 1, \forall 2 \leq \ell \leq L) \\ &\leq \mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_1} = 1)\mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_2} = 1)\mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_\ell} = 1, \forall 3 \leq \ell \leq L) \\ &\leq \Pi_{\ell=1}^{L}\mathbb{P}^{\pi_\theta}(\mathsf{S}_{u_\ell} = 1) \\ &= (1 - \mathrm{e}(\theta))^L.\end{aligned} \tag{42}$$

Since $L \geq \frac{-\log 2}{\log(1 - \mathrm{e}(\theta^* - \delta))}$, it follows that $\mathrm{p}^L(\theta^* - \delta) \leq (1 - \mathrm{e}(\theta^* - \delta))^L \leq 1/2$. Because of the monotonicity of $\mathrm{p}^L(\theta)$, we know that $\theta^* - \delta \leq \theta'$. Moreover, because $\mathrm{p}^L(\theta') = 1/2 > 0$, we have $\theta' < \theta^*$. ■

References

[1] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, Jr., and A. C. Bovik, "A dynamic system model of time-varying subjective quality of video streams over HTTP," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 3602–3606.

[2] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, Jr., and A. C. Bovik, "Modeling the time-varying subjective quality of http video streams with rate adaptations," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2206–2221, May 2013.

[3] "Cisco visual networking index: Global mobile data traffic forecast update, 2011–2016," Cisco, Feb. 2012.

[4] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," in *Proc. ACM SIGCOMM '11 Conf.*, 2011, pp. 362–373.

[5] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

[6] Microsoft Corporation, IIS Smooth Streaming Technical Overview, [Online]. Available: http://www.microsoft.com/en-us/download/default.aspx Sep. 2009

[7] R. Pantos and E. W. May, "HTTP live streaming," *IETF Internet Draft, Work in Progress*, Mar. 2011.

[8] Adobe Systems, "HTTP Dynamics Streaming," Mar. 2012 [Online]. Available: http://www.adobe.com/products/hds-dynamic-streaming.html

[9] MPEG Requirements Group, ISO/IEC FCD Jan. 2011 [Online]. Available: http://mpeg.chiariglione.org/working_documents/ mpeg-b/dash/dash-dis.zip, 23001-6 Part 6: Dynamics adaptive streaming over HTTP (DASH)

[10] C. Luna, Y. Eisenberg, R. Berry, T. Pappas, and A. Katsaggelos, "Joint source coding and data rate adaptation for energy efficient wireless video streaming," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 10, pp. 1710–1720, Oct. 2003.

[11] J. Chakareski and P. Frossard, "Rate-distortion optimized distributed packet scheduling of multiple video streams over shared communication resources," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 207–218, Apr. 2006.

[2] Otherwise, we can simply replace $\mathrm{h}(x)$ with another function whose epigraph is the convex hull of $\mathrm{h}(x)$'s epigraph.

[12] J. Huang, Z. Li, M. Chiang, and A. K. Katsaggelos, "Joint source adaptation and resource allocation for multi-user wireless video streaming," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 5, pp. 582–595, May 2008.

[13] X. Zhu and B. Girod, "Distributed media-aware rate allocation for wireless video streaming," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1462–1474, Nov. 2010.

[14] K. Lin, W.-L. Shen, C.-C. Hsu, and C.-F. Chou, "Quality-differentiated video multicast in multirate wireless networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 1, pp. 21–34, May 2013.

[15] H. Hu, X. Zhu, Y. Wang, R. Pan, J. Zhu, and F. Bonomi, "Proxy-based multi-stream scalable video adaptation over wireless networks using subjective quality and rate models," *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1638–1652, Nov. 2013.

[16] A. Bovik, "Automatic prediction of perceptual image and video quality," *Proc. IEEE*, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[17] M. Barkowsky, B. Eskofier, R. Bitto, J. Bialkowski, and A. Kaup, "Perceptually motivated spatial and temporal integration of pixel based video quality measures," in *Welcome to Mobile Content Quality of Experience*, Mar. 2007, pp. 1–7.

[18] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 253–265, Apr. 2009.

[19] F. Yang, S. Wan, Q. Xie, and H. R. Wu, "No-reference quality assessment for networked video via primary analysis of bit stream," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1544–1554, Nov. 2010.

[20] K. Seshadrinathan and A. C. Bovik, "Temporal hysteresis model of time-varying subjective video quality," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, May 2011, pp. 1153–1156.

[21] J. Park, K. Seshadrinathan, S. Lee, and A. Bovik, "Video quality pooling adaptive to perceptual distortion severity," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 610–620, Feb. 2013.

[22] Y. Huang, S. Mao, and S. Midkiff, "A control-theoretic approach to rate control for streaming videos," *IEEE Trans. Multimedia*, vol. 11, no. 6, pp. 1072–1081, Oct. 2009.

[23] N. Changuel, B. Sayadi, and M. Kieffer, "Control of distributed servers for quality-fair delivery of multiple video streams," in *Proc. 20th ACM Int. Conf. Multimedia*, New York, NY, USA, 2012, pp. 269–278.

[24] C. Yim and A. C. Bovik, "Evaluation of temporal variation of video quality in packet loss networks," *Signal Process: Image Commun.*, vol. 26, no. 1, pp. 24–38, Jan. 2011.

[25] V. Joseph and G. de Veciana, "Jointly optimizing multi-user rate adaptation for video transport over wireless systems: Mean-fairness-variability tradeoffs," in *Proc. IEEE INFOCOM*, 2012, pp. 567–575.

[26] S. Weber and G. de Veciana, "Rate adaptive multimedia streams: Optimization and admission control," *IEEE/ACM Trans. Netw.*, vol. 13, no. 6, pp. 1275–1288, 2005.

[27] J. W. Cho and S. Chong, "Utility max-min flow control using slope-restricted utility functions," *IEEE Trans. Commun.*, vol. 55, no. 5, pp. 963–972, May 2007.

[28] F.-S. Lin, "Optimal real-time admission control algorithms for the video-on-demand (VOD) service," *IEEE Trans. Broadcast*, vol. 44, no. 4, pp. 402–408, Dec. 1998.

[29] P. Mundur, A. Sood, and R. Simon, "Class-based access control for distributed video-on-demand systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 844–853, Jul. 2005.

[30] Y.-H. Tseng, E.-K. Wu, and G.-H. Chen, "An admission control scheme based on online measurement for VBR video streams over wireless home networks," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 470–479, Apr. 2008.

[31] W. Pu, Z. Zou, and C. W. Chen, "Video adaptation proxy for wireless dynamic adaptive streaming over http," in *Proc. 19th Int. Packet Video Workshop (PV)*, 2012, pp. 65–70.

[32] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[33] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[34] *Methodology for the subjective assessment of the quality of television pictures*, ITU-R Rec. BT.500-13, Jan. 2012 [Online]. Available: http://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-I!!PDF-E.pdf,

[35] F. Bellard and M. Niedermayer, FFmpeg, 2012 [Online]. Available: http://ffmpeg.org/

[36] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

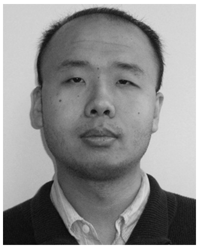
[37] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge Univ. Press, 2004.

[38] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ, USA: Princeton Univ. Press, 1996.

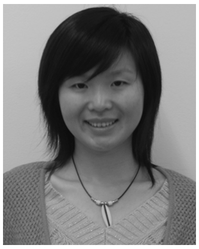
[39] A. L. Stolyar, "Maximizing queueing network utility subject to stability: Greedy primal-dual algorithm," *Queueing Syst.*, vol. 50, no. 4, pp. 401–457, 2005.

[40] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.

[41] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. New York, NY, USA: Springer, 2003.



Chao Chen (S'11–M'14) received the B.E. and M.S. degrees in electrical engineering from Tsinghua University in 2006 and 2009, respectively. In 2009, he joined the Wireless Systems Innovation Laboratory (WSIL) and the Laboratory for Image & Video Engineering (LIVE) at The University of Texas at Austin, where he earned his Ph.D. degree in 2013. Since 2014, he has been working in Qualcomm Incorporated at San Diego. His research interests include visual quality assessment, system identification and network resource allocation.



Xiaoqing Zhu (M'09) is a Technical Leader at the Enterprise Networking Lab at Cisco Systems Inc. She received the B.Eng. degree in electronics engineering from Tsinghua University, Beijing, China. She earned both M.S. and Ph.D. degrees in electrical engineering from Stanford University, California, USA. Prior to joining Cisco, she interned at IBM Almaden Research Center in 2003, and at Sharp Labs of America in 2006. She received the best student paper award in ACM Multimedia 2007.

Dr. Zhu's research interests span across multimedia applications, networking, and wireless communications. She has served as reviewer, TPC member, and special session organizer for various journals, magazines, conferences and workshops. She has contributed as guest editor to several special issues in IEEE Technical Committee on Multimedia Communications (MMTC) E-Letter, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and IEEE TRANSACTIONS ON MULTIMEDIA.

Gustavo de Veciana (S'88–M'94–SM'01–F'09) received his B.S., M.S, and Ph.D. in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993, respectively. He is currently the Joe. J. King Professor at the Department of Electrical and Computer Engineering. He served as the Director and Associate Director of the Wireless Networking and Communications Group (WNCG) at the University of Texas at Austin from 2003–2007.

His research focuses on the analysis and design of wireless and wireline telecommunication networks; architectures and protocols to support sensing and pervasive computing; applied probability and queueing theory. He has served as editor for the IEEE/ACM TRANSACTIONS ON NETWORKING . He was the recipient of a National Science Foundation CAREER Award 1996, co-recipient of the IEEE William McCalla Best ICCAD Paper Award for 2000, co-recipient of the Best Paper in ACM Transactions on Design Automation of Electronic Systems, Jan 2002–2004, co-recipient of the Best Paper in the International Teletraffic Congress (ITC-22) 2010, and of the Best Paper in ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems 2010. In 2009 he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He is on the technical advisory board of IMDEA Networks.

Alan C. Bovik (F'96) is the Curry/Cullen Trust Endowed Chair Professor at The University of Texas at Austin, where he is Director of the Laboratory for Image and Video Engineering (LIVE). He is a faculty member in the Department of Electrical and Computer Engineering and the Center for Perceptual Systems in the Institute for Neuroscience. His research interests include image and video processing, computational vision, and visual perception. He has published more than 650 technical articles in these areas and holds two U.S. patents. His several books include the recent companion volumes The Essential Guides to Image and Video Processing (Academic Press, 2009).

He was named the SPIE/IS&T Imaging Scientist of the Year for 2011. He has also received a number of major awards from the IEEE Signal Processing Society, including: the Best Paper Award (2009); the Education Award (2007); the Technical Achievement Award (2005), and the Meritorious Service Award (1998). He received the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin, the Distinguished Alumni Award from the University of Illinois at Champaign-Urbana (2008), the IEEE Third Millennium Medal (2000) and two journal paper awards from the international Pattern Recognition Society (1988 and 1993). He is a Fellow of the IEEE, a Fellow of the Optical Society of America (OSA), a Fellow of the Society of Photo-Optical and Instrumentation Engineers (SPIE), and a Fellow of the American Institute of Medical and Biomedical Engineering (AIMBE). He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996–1998; co-founder and Editor-in-Chief, IEEE TRANSACTIONS ON IMAGE PROCESSING, 1996–2002; Editorial Board, PROCEEDINGS OF THE IEEE, 1998–2004; Series Editor for Image, Video, and Multimedia Processing, Morgan and Claypool Publishing Company, 2003–present; and Founding General Chairman, First IEEE International Conference on Image Processing, held in Austin, Texas, in November, 1994. He is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.

Robert W. Heath Jr. (S'96–M'01–SM'06–F'11) received the B.S. and M.S. degrees from the University of Virginia, Charlottesville, VA, in 1996 and 1997 respectively, and the Ph.D. from Stanford University, Stanford, CA, in 2002, all in electrical engineering. From 1998 to 2001, he was a Senior Member of the Technical Staff then a Senior Consultant at Iospan Wireless Inc, San Jose, CA, where he worked on the design and implementation of the physical and link layers of the first commercial MIMO-OFDM communication system. Since January 2002, he has been with the Department of Electrical and Computer Engineering at The University of Texas at Austin where he is a Professor and Director of the Wireless Networking and Communications Group. He is also President and CEO of MIMO Wireless Inc. and Chief Innovation Officer at Kuma Signals LLC. His research interests include several aspects of wireless communication and signal processing: limited feedback techniques, multihop networking, multiuser and multicell MIMO, interference alignment, adaptive video transmission, manifold signal processing, and millimeter wave communication techniques.

Dr. Heath has been an Editor for the IEEE TRANSACTIONS ON COMMUNICATION, an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, lead guest editor for an IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS special issue on limited feedback communication, and lead guest editor for an IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING special issue on Heterogenous Networks. He currently serves on the steering committee for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was a member of the Signal Processing for Communications Technical Committee in the IEEE Signal Processing Society. Currently he is the Chair of the IEEE COMSOC Communications Technical Theory Committee. He was a technical co-chair for the 2007 Fall Vehicular Technology Conference, general chair of the 2008 Communication Theory Workshop, general co-chair, technical co-chair and co-organizer of the 2009 IEEE Signal Processing for Wireless Communications Workshop, local co-organizer for the 2009 IEEE CAMSAP Conference, technical co-chair for the 2010 IEEE International Symposium on Information Theory, the technical chair for the 2011 Asilomar Conference on Signals, Systems, and Computers, general chair for the 2013 Asilomar Conference on Signals, Systems, and Computers, general co-chair for the 2013 IEEE GlobalSIP conference, and is technical co-chair for the 2014 IEEE GLOBECOM conference.

Dr. Heath was a co-author of best student paper awards at IEEE VTC 2006 Spring, WPMC 2006, IEEE GLOBECOM 2006, IEEE VTC 2007 Spring, and IEEE RWS 2009, as well as co-recipient of the Grand Prize in the 2008 WinTech WinCool Demo Contest. He was co-recipient of the 2010 *EURASIP Journal on Wireless Communications and Networking* best paper award. He was a 2003 Frontiers in Education New Faculty Fellow. He is the recipient of the David and Doris Lybarger Endowed Faculty Fellowship in Engineering, a licensed Amateur Radio Operator, and is a registered Professional Engineer in Texas.