Abstract—Novel mobility-aware resource allocation schemes have recently been introduced for efficient transmission of stored videos. The essence of such mechanisms is to look-ahead at the future rates users will experience, and then strategically buffer content into user devices when they are at peak radio conditions. For example, a user approaching poor coverage will be preallocated additional video segments to ensure smooth streaming. Advances in mobility prediction and real-time radio environment map updates are driving forces for such Predictive Video Streaming (PVS) mechanisms. Although previous efforts have demonstrated the large potential gains of PVS, ideal channel predictions were assumed. This paper addresses the problem of channel uncertainty in PVS, and proposes a robust resource allocation framework that 1) models channel uncertainty, 2) solves the PVS problem with a tunable level of quality of service guarantees, and 3) learns the degree of uncertainty, and adapts the channel model accordingly. Numerical results demonstrate the effectiveness of the proposed approach for PVS under channel variability.

I. INTRODUCTION

Network operators are facing formidable resource management challenges to cope with the phenomenal growth of mobile traffic. Specifically, video accounted for over 50% of the traffic in 2012, with projections of a 14-fold increase by 2018 [1]. To address this growth, predictive resource allocation techniques that exploit user mobility have been recently proposed to improve throughput and fairness [2], [3], as well as video streaming delivery [4]–[6]. This is accomplished by leveraging the knowledge of the future rates users are expected to experience, and then performing long-term Resource Allocation (RA) plans over several seconds. By doing so, Base Stations (BSs) can schedule more resources to users during their respective peaks, and prioritize users that are headed to poor channel conditions. Such an approach is particularly useful for the delivery of stored videos that can be strategically buffered in advance at the users’ devices. For instance, if it is known a user is approaching a low coverage area, content can be prebuffered to support smooth streaming. Furthermore, as this enables efficient content prebuffering, energy is saved since transmission will not be needed during poor conditions [4]–[6].

The underlying assumption of Predictive Resource Allocation (PRA) approaches is that a user’s future channel states are highly reproducible. This is achieved by coupling mobility information with a Radio Environment Map (REM) or Bandwidth (BW) map, typically generated with road drive tests measuring signal strength and network performance metrics at different locations. Indeed, analyses on human mobility traces reveal that people tend to follow particular routes regularly [7], [8], and several practical studies investigating the correlation between location and received data rates have also been conducted [3], [9], [10]. Yao et al. [9] analyze bandwidth traces collected from two independent cellular providers for routes running through different radio conditions including terrestrial and underwater tunnels. Their findings confirm the correlation between mobile bandwidth and location. The work in [10] conducts a similar measurement study, and addresses other contextual factors such as user speed, time of day, and humidity to predict the available bandwidth more accurately. However, while such maps provide a reasonable estimate of the wireless data rates, they do not accurately capture the dynamics of network congestion or environmental/geographical changes. For instance, a BW map may be more accurate in a rural area and less so in a more urban region. Further, the accuracy of the map may change with time due to fluctuations in network dynamics at rush hours vs. other times. Therefore, there is a need to model rate prediction uncertainty itself, and thereafter develop PRA solutions that incorporate such models. To this end, this paper presents a fuzzy-based robust RA framework Predictive Video Streaming (PVS) under channel uncertainty. We summarize the main contributions of this paper in the following:

• We model uncertainty in the REM measurements using triangular fuzzy numbers. We show that the triangular membership function provides a good approximation of the REM, if the Signal to Noise Ratio (SNR) exhibits a Gaussian prediction error.
• We develop a robust RA framework for predictive video streaming that incorporates the fuzzy REM. The framework allows the operator to control the desired degree of constraint satisfaction (i.e. ensuring video smoothness) under rate uncertainty.
• The proposed framework also ‘learns’ the degree of uncertainty in the REM through feedback and prediction, via a Kalman Filter (KF), and tunes the fuzzy model to reflect the current channel variability.

As opposed to previous works on PRA [2]–[6], we also...

A. Related Work

The work in [4] and [5] are closest to this paper where rate predictions are used to minimize system utilization for stored video delivery. The authors in [4] consider the optimization problem for the multi-user single cell case, and develop optimal RA algorithms for the single user case. In [5], we also discuss the potential energy savings that can be achieved by a mobility-aware wireless access framework. An architecture is presented with the composite functional elements and their interaction is discussed. However, in both these works, ideal channel predictions are assumed and the proposed solutions do not incorporate uncertainty, or provide robust measures to ensure streaming continuity under channel variability. This is addressed in this paper through a Robust Resource Allocation (RRA) framework for PVS. It is worth noting that robust allocation provisions have been proposed for other network functions and application as in [12], [13], where the uncertainty is in the instantaneous Channel Quality Indicator (CQI).

B. Paper Organization

In the following section, we present the system model. Section III presents the predictive streaming optimization problem without channel uncertainty considerations. The proposed RRA framework is then presented in Section IV, and applied to the streaming problem of Section III. We discuss the numerical results in Section V, and conclude in Section VI.

II. SYSTEM MODEL

We use the following notational conventions: $X$ denotes a set and its cardinality is denoted by $|X|$. Matrices are denoted with subscripts, e.g. $x = (x_{a,b} : a \in Z_+, b \in Z_+)$, and $x'$ is the transpose of $x$.

A. System Overview

Consider a BS with an active user set $M$, where an arbitrary user is denoted by $i \in M$. Users enter from the left cell edge and move in a straight line towards the other edge, request stored video content that is transported over HTTP (i.e. as in progressive download). We assume that the wireless link is the bottleneck, and therefore the core network bandwidth is set to 1 Gbps and the video content is always available at the BS.

B. Radio Environment Map and Mobility Information

The REM assumed to be typically available at the service provider would contain the average data rates at different network locations. In order to model such a radio map, we use the Friis Spectrum Loss propagation model in ns-3 [11]. The Signal to Interference plus Noise Ratio (SINR) at each $x$ and $y$ coordinate that users traverse is then computed and the corresponding achievable rate is determined based on the CQI-to-Modulation and Coding Scheme (MCS) mapping in 3rd Generation Partnership Project (3GPP) standards for LTE [14]. We assume that user mobility information is known accurately for the upcoming $T$ seconds, which we call the prediction window, and at a per second granularity. This results in a total of $T$ time slots within the prediction window, which we denote by the set $\mathcal{T} = \{1, 2, \cdots, T\}$. From this information, we construct a matrix of future user rates, defined by $\hat{\mathbf{r}} = (\hat{r}_{i,t} : i \in M, t \in \mathcal{T})$. The values in this matrix will then be fuzzied to account for uncertainty according to the model presented in Section IV-C1.

C. Resource Sharing and Scheduling

BS airtime is shared among the active users during each slot $t$. We define the resource allocation matrix $\mathbf{x} = (x_{i,t} \in [0,1]: i \in M, t \in \mathcal{T})$ which gives the fraction of time during each slot $t$ that the BS bandwidth is assigned to user $i$. The rate received by each user, at each slot, is the element-wise product $\mathbf{x} \odot \hat{\mathbf{r}}$. Airtime sharing is implemented as a time division rate controller over the Round-robin (RR) scheduler in ns-3 [11].

III. PREDICTIVE VIDEO STREAMING: LIMITATIONS OF CRISP RA FORMULATIONS

The essence of predictive video streaming is to strategically transmit content ahead of time at the User Equipment (UE), after which transmission can be momentarily suspended while the user consumes the buffer [4], [5]. If we consider a user requesting a stored video at slot $t = 1$, with a streaming rate of $V$ [bits/s], then the minimum cumulative video content for smooth streaming is $D_{i,1} = V \cdot t$, which is represented with a dashed line in Fig. 1. The cumulative allocation made to a user $i$ by slot $t$ is denoted by $R_{i,t} = \sum_{t'=1}^{t} x_{i,t'} \hat{r}_{i,t'}$. To experience smooth streaming, $R_{i,t} \geq D_{i,t}$ for all $t \in \mathcal{T}$.

Fig. 1 illustrates how BS transmission time can be minimized by leveraging future user rate knowledge. A predictive scheme will wait to make bulk transmissions at times of high channel conditions, while making the minimal transmissions that ensure $R_{i,t} \geq D_{i,t}$ at other times as shown
in Fig. 1(a). This achieves lower airtime usage, resulting in lower power consumption or more resources for other services. The corresponding optimization problem of minimizing BS airtime, without causing any streaming discontinuities can be formulated as the following Linear Program (LP) [5]:

\[
\text{minimize} \quad \sum_{i=1}^{M} \sum_{t=1}^{T} x_{i,t} \quad (1)
\]

subject to: \[
C1: \quad D_{i,t} - R_{i,t} \leq 0, \quad \forall i \in \mathcal{M}, t \in \mathcal{T},
\]
\[
C2: \quad \sum_{i=1}^{M} x_{i,t} \leq 1, \quad \forall t \in \mathcal{T},
\]
\[
C3: \quad x_{i,t} \geq 0 \quad \forall i \in \mathcal{M}, t \in \mathcal{T}.
\]

Constraint C1 ensures that the cumulative video content requirement is not violated at each time slot, which C2 expresses the resource limitation at each base station. It ensures that the sum of the airtime of all users is equal to 1 at every time slot. Finally, C3 provides the bounds for the resource allocation factor.

The solution of the LP in Eq. 1 minimizes airtime without degrading the video only if the predicted rates are accurate. For example, consider the case in Fig. 1(b) where the actual rate is less than the predicted rate. In this case, although airtime is minimized, the user will suffer from video stalls. On the other hand, if the actual rate is greater than the predicted one, then a prebuffering opportunity is lost, resulting in relatively higher total airtime, had the user capitalized on the high rate. To capture and adapt to such variations, we present a fuzzy-based robust RA framework in the following section.

IV. ROBUST RESOURCE ALLOCATION FRAMEWORK FOR PREDICTIVE VIDEO STREAMING

A. Overview

Fig. 2 illustrates the proposed RRA framework for predictive video streaming. The fuzzifier determines the fuzzy rate \( \hat{r} \) which is then used by the rate allocation optimizer to plan the required airtime. The scheduler implements the airtime division among users and measurements of the actual rates experienced \( r_m \) are recorded. This is fed back to a channel variation tracker that predicts the current degree of uncertainty. Based on this, the fuzzifier modifies the membership function of the fuzzy rate \( \hat{r} \) to more accurately reflect the channel variations. The values of \( r_m \) are also fed back to the rate allocator to re-solve the RA problem based on the received user rates. In the following, the RRA framework is presented in detail.

B. Fuzzy-Based Rate Allocation Optimization

The fuzzy rate allocation optimizer used in the RRA framework is based on the fuzzy linear programming model introduced in [15] and [16]. In this approach, the fuzzified rate \( \tilde{r} \) is determined based on 1) the degree of rate uncertainty, and 2) the required level of constraint satisfaction. The formulation in Eq. 1 can be updated to account for the fuzzified rate \( \tilde{r} \) by modifying constraint C1 as follows:

\[
\text{C1 :} \quad D_{i,t} - \sum_{t'=1}^{T} \tilde{r}_{i,t'} x_{i,t'} \leq 0, \quad \forall i \in \mathcal{M}, t \in \mathcal{T}. \quad (2)
\]

Once \( \tilde{r} \) is obtained, an LP solver can be used to solve the airtime minimization problem defined in Eq. 1, with the fuzzy C1 constraint. We now discuss the details of determining \( \tilde{r} \).

C. Fuzzifier: Modeling Rate Uncertainty

1) Rate Membership Function: We represented the fuzzy predicted rate \( \tilde{r}_{i,n} \) by a triangular membership function as shown in Fig. 3. The right \( r_u \) and left \( r_l \) most points on the x-axis define the limits of the triangle’s base, which physically represent the boundaries on the variation of the predicted rate \( \hat{r} \). This can be expressed mathematically as:

\[
\mu_{\tilde{r}} = \begin{cases} 
L(\tilde{r}) = \frac{\tilde{r} - r_l}{r_u - r_l} + 1, & \text{if } r_l \leq \tilde{r} \leq \hat{r} \\
R(\tilde{r}) = \frac{\hat{r} - r_u}{\hat{r} - r_l} + 1, & \text{if } \tilde{r} \leq \hat{r} \leq r_u \\
0, & \text{otherwise}
\end{cases} \quad (3)
\]

This membership function was found to be an acceptable approximation of a Gaussian error overlaid on the predicted rate, as shown in Fig. 4. The step structure appears due to the discrete Adaptive Modulation and Coding (AMC) schemes in LTE, resulting in specific Transport Block (TB) sizes. As some TBs correspond to larger SNR ranges, there are some irregularities in the triangular structure.

2) Defining the Degree of Rate Uncertainty: The \( \alpha \)-cut representation of the membership function indicates the values of fuzzy numbers \( (\tilde{r}_{\alpha,l} \leq \tilde{r} \leq \tilde{r}_{\alpha,u}) \) with a degree of membership that is equal to or greater than \( \alpha \) [15]. The \( \alpha \)-cut of the left side of the triangular membership in Fig. 3 can be determined by setting \( L(\tilde{r}_{\alpha,l}) = \alpha \), and solving for \( \tilde{r}_{\alpha,l} \):

\[
\tilde{r}_{\alpha,l} = \alpha \cdot (\hat{r} - r_l) + r_l. \quad (4)
\]

Similarly, the \( \alpha \)-cut of the right side is:

\[
\tilde{r}_{\alpha,u} = \alpha \cdot (\hat{r} - r_u) + r_u. \quad (5)
\]

Depending on the rate variations (e.g. due to changing radio conditions or network congestion), a suitable value of \( \alpha \) can
be selected to reflect the degree of uncertainty in the predicted rate. For instance, a dynamically changing environment will suffer from wide variations from the predicted rate, and thus a small value of $\alpha$ should be assigned. The corresponding fuzzy rate would include most of the values along the triangle’s base. On the contrary, a higher value of $\alpha$ is suitable for a stable channel, with slight rate variations.

3) Controlling Constraint Satisfaction under Uncertainty: Although each $\alpha$-cut results in a range of possible rates, only one value should be selected (i.e. as the predicted one). This value is then used to solve the rate allocation problem as shown in Fig. 2. This process is performed based on the desired degree of constraint satisfaction, which is denoted by $l_\alpha \in [0, 1]$. A larger $l_\alpha$ corresponds to a higher requirement of constraint satisfaction [15], and therefore a very low rate should be selected from the set of the available rates in the $(\alpha)$-cut. This will result in allocating more airtime to the user to ensure that the streaming constraint Eq. 2 is satisfied even with a pessimistic choice of $\hat{r}$. On the other hand, a higher rate (i.e. an optimistic choice) of $\hat{r}$ can be selected if the constraint satisfaction level is low. This will result in lower airtime and BS resource consumption. In other words, $l_\alpha$ provides an operator trade-off between guaranteeing Quality of Service (QoS) and minimizing BS airtime under rate uncertainty.

4) Determining the Fuzzified Rate ($\tilde{r}$): The degree of rate uncertainty ($\alpha$), and the constraint satisfaction requirement ($l_\alpha$) can be jointly coupled to determine the fuzzified rate ($\tilde{r}$) as illustrated in Fig. 3, and expressed mathematically as [15]:

$$\tilde{r} = l_\alpha \times L(\tilde{r}_{\alpha,l}) + (1 - l_\alpha) \times R(\tilde{r}_{\alpha,u}). \tag{6}$$

Here, the $\alpha$-cut controls the ranges of $L(\tilde{r}_{\alpha,l})$ and $R(\tilde{r}_{\alpha,u})$, while the choice of $l_\alpha$ determines the final value selected within that range. To interpret Eq. 6 further, let us consider the follow cases:

1) **Highest Predictability** ($\alpha = 1$): If $\alpha = 1$, there is no rate uncertainty, and $L(\tilde{r}_{\alpha,l}) = R(\tilde{r}_{\alpha,u}) = \tilde{r}$. Thus, $\tilde{r} = \tilde{r}$, and the optimization problem in Eq. 1 can be solved directly.

2) **Lowest Predictability** ($\alpha = 0$): If $\alpha = 0$, the fuzzy rate will vary between the extreme values of $r_l$ and $r_u$ in Fig. 3. Consequently, depending on the constraint satisfaction $l_\alpha$, the final value of $\tilde{r}$ is determined. For example if:

- $l_\alpha = 0$: Such a choice is suitable if the network operator is either optimistic about the predicted rate, or has a higher preference to efficiency over QoS guarantees. In this case, Eq. 6 reduces to:

$$\tilde{r} = R(\tilde{r}_{\alpha=0,u}) = r_u. \tag{7}$$

- $l_\alpha = 1$: Now the inverse holds, and the operator wants to guarantee constraint satisfaction. According to Eq. 6, lower bound rate is selected:

$$\tilde{r} = L(\tilde{r}_{\alpha=0,l}) = r_l. \tag{8}$$

Note that $l_\alpha$ can vary among users, and provide a mechanism for differentiated QoS. It can also vary with time for the same user, and provide long-term QoS.

D. Adaptive $\alpha$-Tuning: Tracking Rate Variability

E. Adaptive $\alpha$-Tuning: Tracking Rate Variability

In practice, the degree of rate variability will vary with geographical location and time. This lends a fixed value of

![Fig. 3. Triangular membership function of the fuzzy predicted rate $\tilde{r}$ with different $\alpha$-cuts, and $l_\alpha$ values.](image-url)
where $\alpha_t$ is calculated by equating Eq. 6 to $\bar{r}_t$, and then solving for $\alpha$ as depicted below

$$\delta_{\alpha t} = \alpha_t - \alpha_{t-1},$$  \hspace{1cm} (9)

where $\alpha_t$ is calculated by equating Eq. 6 to $\bar{r}_t$, and then solving for $\alpha$ as depicted below

$$\bar{r}_t = r_t - (l_t r_t + (1 - l_t) r_a) / (l_t + (1 - l_t) (r_r - r_u)).$$  \hspace{1cm} (10)

In case of a high variance channel, the value of the measured $\alpha$-cut $\bar{r}_t$ will fluctuate, and thus the error $\delta_{\alpha t}$ will increase. On the other hand, stable channels will result in a fairly equal measured $\alpha$-cut $\bar{r}_t$, and thus the error will start to decrease. Instead of calculating this error based only on the current measurements, a Kalman Filter (KF) will be used to track the error while considering its previous values as well. KFs have been used widely in several applications to provide an optimal estimation of dynamically changing systems [17]. The KF assumes that the error estimation is a Gaussian linear stochastic process in discrete time, and thus the measurement and system noises are uncorrelated random processes with a normal distribution and zero mean. Accordingly, if the rate error follows a log-normal function, it is suitable to apply the KF to predict the rate variability. The standard KF operations and equations are summarized below [17]:

**Prediction Phase:**

$$X^- \approx \Phi X^-+$$  \hspace{1cm} (11)

$$P^-_t = \Phi P^-+_{t-1} \Phi^T + Q.$$  \hspace{1cm} (12)

**Measurement Phase:**

$$K_t = P^- (H_t P^- H^T + R)^{-1}$$  \hspace{1cm} (13)

$$X^+_t = X^- + K_t(z_t - H_t X^-_t)$$  \hspace{1cm} (14)

$$P^+_t = P^- - K_t H_t P^-$$  \hspace{1cm} (15)

where $X^+_t$ and $X^-_t$ are the priori and posterior error values respectively. $P^-_t$ and $P^+_t$ are the error estimation matrices respectively. $H$ and $\Phi$ are the observation and state transition matrices respectively, while $Q$ and $R$ are the process and the measurement noise covariance matrices respectively, and $K$ is the Kalman filter gain.

The Kalman filter operates in two consecutive stages: Prediction and Measurement. In the former one, the predicted state value $X^-_t$ is calculated based on its value in the previous time slot $X^-_{t-1}$ as in Eq. 11. In the measurement phase, the observation $z_t$ is used to modify the predicted state $X^-_t$ based on the current Kalman filter gain $K_t$, resulting in the calculation of the new state value $X^+_t$ as in Eq. 13-14. In our model, the predicted state $X^-_t$ represents the error in the calculated degree of uncertainty $\delta_{\alpha t}$ and is assumed to be the same as the corrected error of the previous time step $X^-_{t-1}$. Thus, the state transition matrix is set to unity. The observation $z_t$ represents the current error in the degree of uncertainty based on the current measurements $\delta_{\alpha t}$ shown in Eq. 9. The observation is updated every time slot based on the average measured rate $\bar{r}_t$. Since the observations $z_t$ and the predicted state value $X^-_t$ represent values for the errors in the degree of uncertainty, the state observation matrix $H$ is set to unity. The values of $Q$, $R$ and the initial value of $P$ ($P^0_t$) are obtained from excessive tuning and their values are shown in Table I.

In summary, the KF standard equations are modified for our problem as follows:

**Prediction Phase:**

$$\delta_{\alpha t} = \delta_{\alpha t-1}$$  \hspace{1cm} (16)

$$P^-_t = P^-+_{t-1} + Q.$$  \hspace{1cm} (17)

**Measurement Phase:**

$$K_t = P^- (H_t P^- H^T + R)^{-1}$$  \hspace{1cm} (18)

$$\delta_{\alpha t} = \delta_{\alpha t} + K_t(\delta_{\alpha t} - \delta_{\alpha t})$$  \hspace{1cm} (19)

$$P^+_t = P^- - K_t H_t P^-$$  \hspace{1cm} (20)

2) $\alpha$-Tuning Utility: When rate variations are low, $\delta_{\alpha} \approx 0$, which should correspond to $\alpha \approx 1$. Contrarily high variations result in larger values of $\delta_{\alpha}$, which should lead to $\alpha \approx 0$. This mapping is accomplished using the following utility function

$$\alpha = 1 - e^{-\gamma |\delta_{\alpha}|},$$  \hspace{1cm} (21)

where $\gamma$ controls the rate of utility decrease with increasing $\delta_{\alpha}$. Other sigmoid-like utilities may also be used to accomplish this mapping.
V. PERFORMANCE EVALUATION

A. Simulation Set-up

The simulation is performed using the LTE module in the Network Simulator (ns-3) [11], with model parameters as indicated in Table I. Gurobi [18] is used to solve the rate allocation optimization and is integrated in the simulator. Average BS airtime and video degradation (VD) are used as performance metrics. Video degradation is defined as the total fraction of constraint violation, i.e. the constraint in Eq. 2. Note the gains of the predictive RA over traditional scheduling approaches has been reported in [4]–[6], so we therefore focus on studying the proposed RRA under uncertainty.

B. Effect of Constraint Satisfaction (lα)

In Fig. 5, we investigate the effect of the constraint satisfaction level lα, for a single user crossing the cell. The results were averaged over 100 different log-normal error distributions for different variances σ, and predictability levels (α-cuts). We can make the following observations:

- Fig. 5(a) shows that as lα increases, the consumed airtime increases. This is expected since higher lα values will result in a smaller r̂, thereby requiring more airtime to satisfy the constraint in Eq. 2. The result is lower VD as illustrated in Fig. 5(b), since the constraints are satisfied with higher probabilities.

- To almost eliminate VD, lα ≈ 1 and α ≈ 0.25. This cause a sharp increase in the consumed airtime.

- As the α-cuts increase, the airtime increases for lα < 0.5, but decreases for lα > 0.5. The reason is that lα ≈ 0.5 is an inflection point, where r̂ > r̂ for lα < 0.5, while r̂ < r̂ for lα > 0.5 as illustrated in Fig. 3. As α decreases, the deviation from r̂ increases, so the effects are more pronounced at both extremes of lα. A similar reasoning can be applied to trends of VD in Fig. 5(b).

- Average airtime is less for higher error variances σ. Referring to Fig. 4, we can see that a higher variance has a higher probability or large TB transmissions. Even through there is a higher probability of small TB transmissions as well, the overall effect is a reduced airtime since a few large TBs are sufficient to buffer the video. As a result, the VD is generally less for higher σ. However, this is not the case for larger degrees of constraint satisfaction lα, where it is paramount that the perceived user rates are greater than the value of r̂. With a high variance this is not guaranteed as illustrated in Fig. 4.

C. Effect of Rate Variations and Adaptive α-Tuning

We now investigate the effect of variable degrees of channel uncertainty and the potential gains or adapting α with time. In this scenario, the error variance is initially high σ = 6, and then decreases to σ = 2 as the user approaches the cell center. The airtime and VD results are jointly plotted in Fig. 6 for different values of α-cuts (at an lα = 0.75), and compared to the adaptive-α scheme based on the KF. We can see that lowering α reduces degradation, but at the cost of an increased airtime. However, keeping α constantly low, is not ideal in this case since the channel error variance decreases to σ = 2 during the simulation. The proposed adaptive-α approach is able to decrease its value when the channel variance is high in order to avoid VD, and then increase α when the channel variance is low in order to satisfy the constraint with lower airtime. This results in an acceptable VD with a low airtime as illustrated in Fig. 6.

This scenario was extended to the multi-user case, where users enter the cell with an inter-arrival rate of 5 seconds. The results in Fig. 7 further emphasize the importance of adapting α. As the number of users increase, the competition on the radio resources increases and thus accurate adaptation yields significant VD improvements without compromising total airtime as shown in Fig. 7(a). This indicates that the design of KF successfully learned the channel variation and tuned α to satisfy the constraints without resource wastage.

VI. CONCLUSION

The potential of exploiting rate predictions to transmit stored videos efficiently has been recently investigated with promising results [4]–[6]. This is based on numerous human mobility studies and analyses of BW traces demonstrating that a user’s future channel states are highly reproducible [7]–[10]. While these studies indicate the importance of predictive RA techniques, ideal channel predictions were assumed. A study is therefore needed to investigate and design robust RA for PVS under channel uncertainty. To this end, we developed a fuzzy-based RRA framework that incorporates channel uncertainty into the PVS problem, and provides a tunable level of service guarantees. We also find that it is important to learn the degree of uncertainty in order to meet the desired constraint satisfaction levels without unnecessary resource consumption. To this end, we incorporate feedback in the framework to learn and adapt to the degree of channel uncertainty, and re-optimize the RA in PVS. A detailed numerical analysis of the framework was conducted to investigate the effects of channel variability and provide insights to further developments. Future work includes studying the performance in more complex simulation settings, as well as investigating the use of stochastic models and other channel variability predictors as alternative approaches to be used within the RRA framework.

REFERENCES

Fig. 5. Video degradation fraction VD and average BS airtime for varying constraint satisfaction levels $l_{\alpha}$, $\alpha$-cuts, and error variances $\sigma$.

Fig. 6. Effect of adapting $\alpha$, with $l_{\alpha} = 0.75$ with multi-variance error.
Fig. 7. Effect of adapting $\alpha$, with $l_\alpha = 0.75$ in the multi-user, multi-variance error scenario.