Chance-Constrained QoS Satisfaction for Predictive Video Streaming

Ramy Atawia*, Hatem Abou-zeid*, Hossam S. Hassanein† and Aboelmagd Noureldin‡
*Electrical and Computer Eng. Dept., Queen's University, Canada, {ramy.atawia, h.abouzeid}@queensu.ca
†School of Computing, Queen's University, Canada, hossam@cs.queensu.ca
‡Electrical and Computer Eng. Dept., Royal Military College of Canada, Canada, aboelmagd.noureldin@rmc.ca

***Abstract*—The promising energy saving and QoS gains of Predictive Resource Allocation (PRA) techniques have recently been recognized in the wireless network research community. These gains were primarily introduced in light of perfect prediction of both mobility traces and anticipated channel rates. However, under real world considerations of prediction errors, the reported gains cannot be guaranteed and further investigation is needed. In this paper, we demonstrate the practical potential of PRA by developing a robust, probabilistic framework that guarantees QoS satisfaction for video streaming under imperfect predictions, without compromising the energy saving gains. The proposed PRA framework uses chance-constrained programming to model video streaming QoS for all users during the foreseen time horizon. Closed form solutions are developed using the Gaussian and Bernstein approximations based on the channel statistical measures. Extensive numerical simulations using a standard compliant Long Term Evolution (LTE) system are presented to examine the developed solutions, for different user mobility scenarios and target QoS levels. The results demonstrate the various design trade-offs involved toward the practical deployment of predictive video streaming in future generation networks.**

I. Introduction

Energy saving in wireless networks is receiving growing interest from operators to increase revenues while minimizing environmental impact. In particular, novel *energy-efficient* Radio Access Network (RAN) frameworks are urgently needed since RANs account for more than half the network energy consumption [1]. This is further accentuated with the predictions that mobile data traffic will experience a compound annual growth rate (CAGR) of 57% from 2014 to 2019, reaching 24.3 exabytes per month by 2019. Approximately 70% of this traffic, or 17.4 exabytes of monthly data will be mobile video streaming. To cope with these challenges, novel wireless resource allocation schemes targeting *video delivery* are of paramount importance.

Recent positioning techniques enable the high predictability of users' trajectories in different environments [2], [3]. In addition, studies on human mobility indicate that people tend to follow the same movement patterns [4], [5], and thus, the signal strength levels of mobile users can be highly predictable [6], [7]. Consequently, Predictive Resource Allocation (PRA) that leverages user signal strength patterns over a time horizon, has recently been proposed to improve video streaming quality [8], [9], and reduce transmission energy [10], [11]. The essence of PRA is to apply *long-term* allocation plans over several seconds, by exploiting knowledge of the *future* user link capacities. This enables Base Stations (BSs) to prioritize users moving towards poor radio conditions, or delay transmission until a user reaches better channel conditions. Stored video content such as YouTube and Netflix is well suited for such approaches as it can be strategically prebuffered and stored on the local cache of the User Equipment (UE).

The potential gains of PRA recently reported in literature [8]–[13] are very encouraging, and demonstrate the need for further investigation. In particular, *ideal* predictions of future data rate with deterministic Quality of Service (QoS) constraints were primarily used. With such approaches, evaluating system performance under real world uncertainty is challenging, and probabilistic QoS *guarantees* are not possible. Furthermore, measurement studies reveal that the predictability of signal strength varies significantly with geographical location and time of day [7]. Stochastic PRA approaches are therefore needed to 1) model the rate uncertainty adaptively, and 2) incorporate probabilistic models that can strike a balance between providing high gains when predictions are accurate, and minimize the risks associated with erroneous predictions during periods of uncertainty.

For the first time in literature, this paper introduces a *stochastic*, predictive RAN framework that minimizes BS energy consumption for video delivery to mobile users. This is achieved while providing *robust* QoS guarantees that can be tuned based on user priority and operator objectives. Herein, we model the desired QoS satisfaction level as a probabilistic chance constraint in which predicted rates are random variables, rather than the expected values used in previous works. Two tractable deterministic forms, namely the *Gaussian* and *Bernstein*, are then adopted to attain a closed form solution for the probabilistic formulation. The first method requires an invertible Cumulative Density Function (CDF) of the predicted rate, while the latter is an alternative that requires only rate bounds, and thus can consider more general prediction errors. The resulting energy savings and QoS satisfaction levels are then assessed and compared to evaluate their effective usage within the proposed framework. We believe this work provides a direction toward the development of practically deployable predictive RANs in future generation networks.

In the following section, we provide a background to Chance Constrained Programming (CCP), and review the related literature. Section III presents the PRA problem definition and highlights the limitations of the non-robust PRA. The proposed CCP based PRA framework for energy-efficient

978-1-4673-6770-7/15/$31.00 ©2015 IEEE

video streaming is then presented in Section IV. Simulation results are discussed in Section V and finally, we conclude the paper in Section VI.

II. Background and Related Work

Variations in received signal strengths over wireless channels are typically modeled as random variables. Conventional Resource Allocation (RA) typically represent this randomness by the expected value in the optimization problem. However, the resulting deterministic formulation does not guarantee QoS satisfaction as users may experience lower data rates than the utilized expected values.

Robust stochastic optimization techniques represent the random variable by its statistical properties (probability density function, variance and mean) rather than the mean value only [14]. The formulated problem comprises probabilistic chance constraints that can guarantee QoS satisfaction by a certain predetermined level $1-\epsilon \in [0,1]$. Stochastic optimization can also consider randomness in the objective function coefficient as well [15]. In this paper, we focus on randomness in constraint coefficients which are handled by CCP, initially introduced in [16]. Essentially, the general formulation of CCP for QoS satisfaction in RA is shown in Eq. 1 as follows:

$$Pr\left\{f(x_t,\eta_t) \geq D_t\right\} \geq 1-\epsilon, \qquad \forall t \in \mathcal{T}, \tag{1}$$

where $f(x_t,\eta_t)$ is the RA function at time slot t which incorporates the decision variable x_t, and the random rate η_t. D_t is the user demand at time slot t and $\mathcal{T}$ is the allocation decision time horizon. The above CCP formulation ensures the satisfaction of demand D_t at each allocation time slot t with a minimum probability of $1-\epsilon$. In other words, the probability of violating the demand and therefore the QoS is bounded by ϵ. CPP has been widely applied in several non-predictive based optimizations for wireless resource allocation such as OFDM scheduling [17] and channel assignment [18].

The main challenge of CCP is that the resulting problem is no longer solvable by traditional mathematical optimization techniques. Simulation based and analytical approaches have therefore been introduced as solution mechanisms. Examples of simulation based methods include realtime Monte-Carlo simulations [14] where a large number of samples for the random variable are drawn according to its probability density function (PDF). Thus, the CCP is converted to a number of non-probabilistic formulations each solved individually. Then, the optimal solution is determined to be the one that satisfies $(1-\eta)\times 100\%$ of the generated scenarios. However, the main disadvantage of such methods is that optimality depends on the number of samples [19] especially in high QoS levels to accurately approximate the original PDF. Thus, this approach is not easily applicable for realtime allocations. Alternatively, analytical approaches such as Gaussian and Bernstein approximations [19] provide a deterministic closed form for the CCP that can be solved with mathematical optimization techniques. In particular, the Gaussian approximation (GA) exploits the CDF and its inverse in deriving the deterministic form. For cases of non-invertible or costly CDF computation, the Bernstein approximation (BA) can be utilized. Herein, the BA uses the moment generating function and its cumulant to develop the deterministic equivalent of the CCP. Further approximations [20] can be also applied to represent the final deterministic form using only the maximum bounds of the random rate.

In our *energy-efficient robust* PRA framework, both GA and BA are applied. The former fits the recent measurement for modelling rate prediction errors as a family of Gaussian processes [21]. On the other hand, the BA approach is considered to extend the GA for general types of uncertainties. This work differs from previous predictive resource allocation studies that did not incorporate uncertainty and probabilistic QoS guarantees [8]–[11], [13], and our work in [22] that only considered fuzzy-based uncertainty in predicted user rates.

III. System Overview

A. Preliminaries

We use the following notational conventions throughout the paper: $\mathcal{X}$ denotes a set and its cardinality is denoted by X. Matrices are denoted with subscripts, e.g. $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$. $\tilde{r}$ represents a normal random variable (r.v.), and its cumulative density function is denoted as Q. The inverse Q function of a zero mean and unit variance normal r.v. is denoted by Q^{-1}. $\mathbb{E}[\cdot]$ denotes the expectation of a r.v. and $log(\cdot)$ denotes the natural logarithmic function.

B. Problem Definition and Limitations of Non-Robust PRA

The system considers a BS with an active user set $\mathcal{M}$ where the user index is denoted by $i \in \mathcal{M}$. Each mobile user requests video with a fixed streaming rate. The video is then transmitted from the server to the Evolved Packet Core (EPC) and then to the BS. We assume that user's mobility trace is known for the next T seconds, called the prediction window, and at a per second granularity[1]. This results in a total of T time slots within the prediction window, which we denote by the set $\mathcal{T} = \{1, 2, \cdots, T\}$. The active users share the BS resources (airtime fractions) at each time slot t. The resource allocation matrix $\mathbf{x} = (x_{i,t} \in [0,1] : i \in \mathcal{M}, t \in \mathcal{T})$ gives the fraction of time slot t during which BS's bandwidth is assigned to user i. The average available rate for user i at time slot t is denoted as $\bar{r}_{i,t}$, which is calculated by mapping the predicted user traces to the Radio Environment Map (REM) at the service provider.

The PRA problem addressed in this paper aims to satisfy the users QoS level while minimizing the total BS airtime User's QoS is said to be satisfied when the video is played back smoothly without stops. Quantitatively, this is achieved when total amount of data delivered to the user at a certain time slot is not less than the *cumulative* demanded data. The above predictive green video transmission strategy was introduced in [11], [12] and can be formulated as follows

[1]Note that in this paper our focus is on modeling the uncertainty in the rate prediction, and designing a robust PRA, but not on user mobility prediction itself.

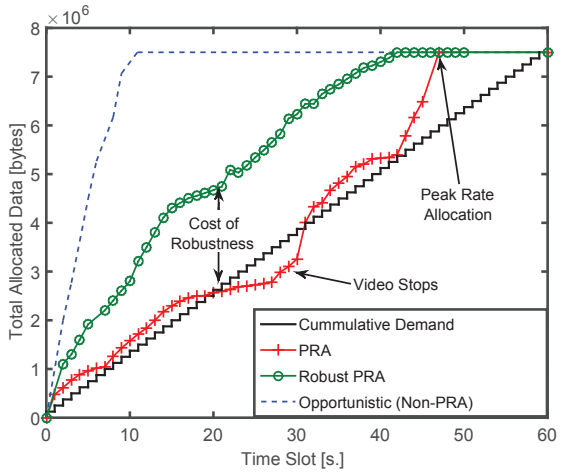


Fig. 1. Examples of resource allocation strategies.

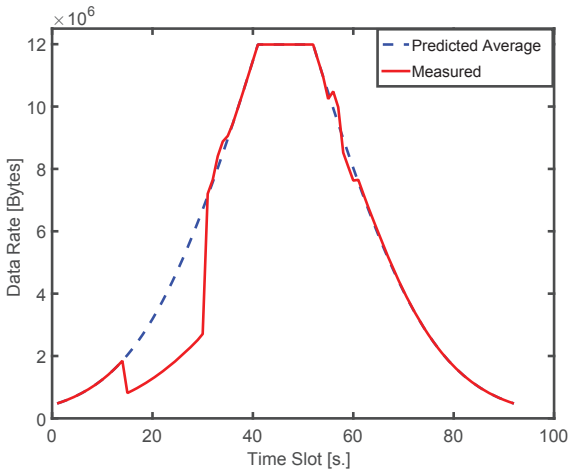


Fig. 2. Examples of predicted rate variations.

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{M} x_{i,t} \tag{2}$$

subject to:

$$\text{C1: } \sum_{t=0}^{t'} \bar{r}_{i,t}x_{i,t} \geq D_{i,t'} \quad ,\forall\, i \in \mathcal{M},\ t' \in \mathcal{T},$$

$$\text{C2: } \sum_{i=1}^{M} x_{i,t} \leq 1, \qquad \forall t \in \mathcal{T},$$

$$\text{C3: } x_{i,t} \geq 0 \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

The QoS constraint C1 ensures the smooth playback of the video as the cumulative sent video content must be greater than the cumulative demanded streaming data $D_{i,t'}$. The second constraint models the limited resources at each base station by ensuring that the sum of the airtime of all users is less than 1 at every time slot. Finally, C3 ensures the non-negativity of the assigned airtime fractions. The above allocation was shown to provide both energy savings and QoS satisfaction under perfect future channel knowledge [10], [11].

Unlike traditional RA, the idea of energy-efficient PRA is to wait until the user reaches peak radio conditions, and then push large portions of the video to avoid future allocation in the lower data rates. Before the user reaches the peak channel conditions, QoS is met by allocating the *minimum* amount of data to the user, i.e., the total allocated data is maintained to be exactly equal to the demand before the peak is encountered. An example of such an energy saving allocation strategy is illustrated in Fig. 1. However, and as discussed previously, the above formulation depends on the *average* value of future data rates and thus it is not robust to any channel variations. QoS satisfaction is therefore not guaranteed when lower rates are experienced by the user due to imperfect predictions. This limitation is illustrated in Fig. 1 where the QoS is not met in the interval between 20 s and 30 s as the actual rate dropped below its average predicted value as illustrated in Fig. 2.

IV. Chance Constrained Problem Formulation

The robust PRA framework for video streaming uses probabilistic chance constraints to model the QoS guarantees. This is achieved by replacing the deterministic QoS constraint C1 in Eq. 2 with the probabilistic chance form, where the expected values of the predicted rates are replaced by random variables as follows:

$$\underset{\mathbf{x}}{\text{minimize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{M} x_{i,t} \tag{3}$$

subject to:

$$\text{C1: } Pr\left\{\sum_{t=0}^{t'} \tilde{r}_{i,t}x_{i,t} \geq D_{i,t'}\right\} \geq 1-\epsilon, \qquad \forall\, i \in \mathcal{M},\ t' \in \mathcal{T},$$

$$\text{C2: } \sum_{i=1}^{M} x_{i,t} \leq 1, \qquad \forall t \in \mathcal{T},$$

$$\text{C3: } x_{i,t} \geq 0 \qquad \forall\, i \in \mathcal{M}, t \in \mathcal{T}.$$

where $\epsilon \in [0, 1]$ is the maximum probability that video stops occur, and takes values less than 0.5 for reliable performance.

In order to obtain a closed form solution of Eq. 3, the probabilistic constraint should be replaced with a tractable deterministic equivalent form. Generally in this form, the random rate is replaced by its expected value $\bar{r}_{i,t}$ in addition to a subtracted safety term S. This safety term is a function of: the statistical parameters (e.g. variance σ^2) of the predicted random rate, the QoS level $(1-\epsilon)$ and the decision variable $x_{i,t}$. The resulting L.H.S should remain to exceed the demand under variations in the predicted rate as depicted below:

$$\sum_{t=0}^{t'} \bar{r}_{i,t}x_{i,t} - S \geq D_{i,t'}, \tag{4}$$

where $S = f(\sigma^2_{i,t}, \epsilon, x_{i,t})$. The value of the safety function is directly proportional to both the degree of constraint satisfaction $(1-\epsilon)$ and the variance $\sigma^2_{i,t}$ of the predicted rate.

For instance, for high values of $(1-\epsilon)$, the safety term should increase in order to decrease the total equivalent predicted rate. Thus, more airtime (higher $x_{i,t}$) is allocated to compensate the reduced rate and ensure that the demand $D_{i,t}$ is met as illustrated in Fig. 1. Similarly, the safety term should be increased to account for the scenarios where the predicted rate drops dramatically below its expected value (i.e., high $\sigma^2_{i,t}$).

Deriving the safety term should consider both the optimality and feasibility of the obtained allocation. Highly conservative safety terms guarantee QoS satisfaction at a degree much more than $(1-\epsilon)$ even at higher variations. However, this occurs at the expense of increased airtime, which decreases the energy saving gain of the PRA. On the other hand, optimistic less conservative safety terms allocate less airtime, and thus have a high risk of violating the QoS level. In this paper, we discuss two commonly used approaches to develop the safety term in Eq. 4, namely the Gaussian and Bernstein approximations. Generally, to derive the deterministic equivalent and the safety term, the Gaussian approach requires an invertible CDF for the predicted random rate. On the other hand, the Bernstein approximation requires only the bounds of the predicted rate. A background on the methodology and assumptions of both methods is also provided, and the degrees of conservatism of both techniques are then discussed and compared.

A. Gaussian Approximation

In GA, the predicted channel rate uncertainty is a normally distributed random variable denoted as: $\tilde{r}_{i,t'} \sim N(\bar{r}_{i,t}, \sigma^2_{i,t})$. The summation of $\tilde{r}_{i,t'}$ in C1 of Eq. 3 results in a multivariate normally distributed variable with mean μ and variance-covariance matrix Σ as depicted below:

$$\mu_i = \sum_{t=0}^{t'} \bar{r}_{i,t}, \qquad \forall i \in \mathcal{I}, \forall t' \in \mathcal{T} \tag{5}$$

$$\Sigma = \begin{bmatrix} {\sigma^r_{i,0}}^2 & \dots & \sigma^r_{i,0,t'} \\ \dots & {\sigma^r_{i,1}}^2 & \dots \\ \sigma^r_{i,t',0} & \dots & {\sigma^r_{i,t'}}^2 \end{bmatrix}, \quad \forall i \in \mathcal{I}, \forall t' \in \mathcal{T} \tag{6}$$

Where: $\bar{r}_{i,t} = \mathbb{E}[\tilde{r}_{i,t}]$, $\sigma^r_{i,t,k} = \mathbb{E}[(\tilde{r}_{i,t} - \bar{r}_{i,t})(\tilde{r}_{i,k} - \bar{r}_{i,k})]$ is the covariance of both rates $\tilde{r}_{i,t}$ and $\tilde{r}_{i,k}$. ${\sigma^r_{i,t}}^2 = \mathbb{E}[(\tilde{r}_{i,t} - \bar{r}_{i,t})^2]$ is the variance of rate $\tilde{r}_{i,t}$.

The deterministic closed form of Eq. 3 can be expressed using the multivariate random variables and normal CDF as follows:

$$\begin{aligned} Q\left(\frac{D_{i,t'} - \sum_{t=0}^{t'} \bar{r}_{i,t} x_{i,t}}{\sqrt{\sum_{t=0}^{t'} \sum_{k=0}^{t'} x^2_{i,t} \sigma^r_{i,t,k}}}\right) \geq 1-\epsilon, \\ \forall i \in \mathcal{M}, t' \in \mathcal{T}, \\ \sum_{t=0}^{t'} \bar{r}_{i,t} x_{i,t} - Q^{-1}_{\epsilon} \sqrt{\sum_{t=0}^{t'} \sum_{k=0}^{t'} (x_{i,t} \sigma^r_{i,t,k})^2} \geq D_{i,t'}, \\ \forall i \in \mathcal{M},\ t' \in \mathcal{T}. \end{aligned} \tag{7}$$

Assuming that the random rates experienced by each user over the time slots are independent, $\sigma^r_{i,t,k} = 0, \forall t \neq k$, and Eq. 7 reduces to Eq. 8 below:

$$\sum_{t=0}^{t'} \bar{r}_{i,t} x_{i,t} - Q^{-1}_{\epsilon} \sqrt{\sum_{t=0}^{t'} (x_{i,t} \sigma^r_{i,t})^2} \geq D_{i,t}, \qquad \forall i \in \mathcal{M},\ t' \in \mathcal{T}. \tag{8}$$

The above constraint representation is a second order cone programming (SoCP) model whose convexity is guaranteed for $\epsilon < 0.5$ and $x_{i,t} \in [0,1]$.

B. Bernstein Approximation

Unlike GA, Bernstein's approximation utilizes the marginal distribution and the moment generating function instead of the inverse CDF. BA represents the chance constraint as a linear summation of random variables as follows

$$Pr\left(f_0(\mathbf{x}) + \sum_{t=1}^{t'} \eta_t f_t(\mathbf{x}) \leq 0\right) \geq 1-\epsilon, \qquad \forall t' \in \mathcal{T}. \tag{9}$$

Here η_t is the random variable with marginal distribution $\mathbb{P}_t$, and $f_t(\mathbf{x})$ is a convex function containing the decision vector $\mathbf{x}$. Assuming that all the random variables η_t are independent, $\mathbb{P}_t$ has a bounded support on the interval $[-1,1]$ $\forall t$ and the function $f_t(\mathbf{x})$ is affine in the decision vector x. Therefore, with the aforementioned assumptions, a convex deterministic equivalent for Eq. 9 can be obtained as follows

$$\inf_{\lambda > 0} \left[f_0(\mathbf{x}) + \sum_{t=1}^{t'} \lambda \Lambda_t(\lambda^{-1} f_t(\mathbf{x})) + \lambda \log \frac{1}{\epsilon} \right] \leq 0, \qquad \forall t \in \mathcal{T}. \tag{10}$$

Herein, $\Lambda_t(z)$ is the logarithm of the moment generating function $M_t(z)$ for r.v. z as depicted in Eq. 11

$$\begin{aligned} \Lambda_t(z) &= log M_t(z) \\ M_t(z) &= \mathbb{E}\left[e^{kz}\right] = \int e^{kz} d\mathbb{P}_t(k) \end{aligned} \tag{11}$$

Instead of computing the exact value of the logarithm moment generating function in Eq. 11, in addition to solving for the auxiliary variable λ, a conservative approximation using the upper bound can be adopted as in Eq. 12 [20].

$$\begin{aligned} \Lambda_t(z) \leq max\left\{\mu^+_t z, \mu^-_t z\right\} + \frac{\sigma^2_t}{2} z^2, \forall t \in \mathcal{T} \\ -1 \leq \mu^-_t \leq \mu^+_t \leq 1 \end{aligned} \tag{12}$$

The variables μ^+_t, μ^-_t and σ_t are used to approximate the bounded support and their corresponding numeric values are available in [20]. Therefore, a conservative deterministic equivalent for Eq. 10 is attained using Eq. 12 and the arithmetic inequality as follows

$$f_0(\mathbf{x}) + \sum_{t=1}^{t'} max\left\{\mu_t^+ f_t(\mathbf{x}), \mu_t^- f_t(\mathbf{x})\right\} + \sqrt{2log(\frac{1}{\epsilon})\left(\sum_{t=1}^{t'} \sigma_t^2 f_t(\mathbf{x})^2\right)} \leq 0, \qquad \forall t' \in \mathcal{T}. \tag{13}$$

Finally, the robust PRA chance constraint C1 in Eq. 3 is replaced by Eq. 13 as depicted in Eq. 14:

$$\sum_{t=1}^{t'} \bar{r}_{i,t} x_{i,t} + \sum_{t=1}^{t'} \mu_{i,t}^- \hat{r}_{i,t} x_{i,t} - \sqrt{2log(\frac{1}{\epsilon})\left(\sum_{t=1}^{t'} (\sigma_{i,t} \hat{r}_{i,t} x_{i,t})^2\right)} \geq D_{i,t'}, \quad \forall t' \in \mathcal{T}, \tag{14}$$

where the random predicted rate $\tilde{r}_{i,t}$ is assumed bounded in $[r_{i,t}^l, r_{i,t}^u]$. To satisfy the assumptions for Eq. 10, this rate is normalized in $[-1, 1]$ by using the maximum deviation and the average values denoted by $\hat{r}_{i,t}$ and $\bar{r}_{i,t}$ respectively:

$$\hat{r}_{i,t} = \frac{r_{i,t}^u - r_{i,t}^l}{2}, \quad r_{i,t}^u > r_{i,t}^l$$
$$\bar{r}_{i,t} = \frac{r_{i,t}^u + r_{i,t}^l}{2} \tag{15}$$

Similar to the GA, the above constraint is also an SoCP model which is convex for $\epsilon < 0.5$ and $x_{i,t} \in [0, 1]$.

C. Monte Carlo Statistical Parameters Estimation

In order to obtain optimal values of safety terms, the statistical measures of the rate (i.e., ${\sigma_{i,t}^r}^2$ and $\hat{r}_{i,t}$) need to be determined. Lower values of ${\sigma_{i,t}^r}^2$ or $\hat{r}_{i,t}$ than the actual measurements will result in a small value of the safety term which increases the risk violating the QoS level, and the converse is true. To address this, off-line Monte Carlo simulations are adopted prior to solving the RA problem. The simulation generates all the possible channel rates and adds random errors to them to build the rate distribution function.

Different values of the signal to interference plus noise ratio (SINR) are generated. For each value, the corresponding rate is calculated and denoted as R. Concurrently, N random samples are generated and added to the current SINR to generate erroneous SINR denoted as $SINR_e$. Then, N rates are constructed from $SINR_e$ and denoted as R_e. These rates are used to construct the probability distribution $\mathbb{P}$ of rate R. The simulation continues to generates a new value of SINR and repeats the above procedure until the maximum rate is generated. Finally, the bounds of each distribution and the variance are calculated while considering R to be the mean value. It is worth noting that the SINR is mapped to the corresponding CQI level using formulas in [23]. The latter is then converted to the channel rate using the bandwidth (BW) and bit error rate (BER) values according to [24], and the generated error follows the 3GPP correlated fading model in [25]. All the above steps are summarized in Fig. 3. The main advantage of performing the above estimation off-line is to generate large samples of both the SINR and the added random variables. This results in highly accurate statistical estimation of the parameters used in the robust PRA.

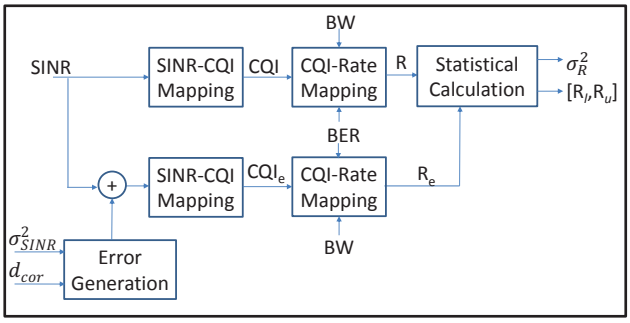


Fig. 3. Block diagram for generating statistical parameters of the predicted rates using offline Monte-Carlo simulations

TABLE I
SUMMARY OF MODEL PARAMETERS

Parameter	Value
BS transmit power	43 dBm
BW	5 MHz
T	60 s
Streaming Rate	1 [Mbps]
BER	5×10^{-5}
Shadow correlation distance (m)	37 (LoS) and 50 (NLoS)
Shadow standard deviation (dB)	3 (LoS) and 6 (NLoS)
Velocity (km/h)	30 (Urban) 65 (Rural)
μ^-	-0.5
σ_t'	$\frac{1}{\sqrt{12}}$
Packet size	10^3 [bytes]
Packet rate (EPC to BS)	$10^3 s^{-1}$
Total number of packets	7.5×10^3

V. PERFORMANCE EVALUATION

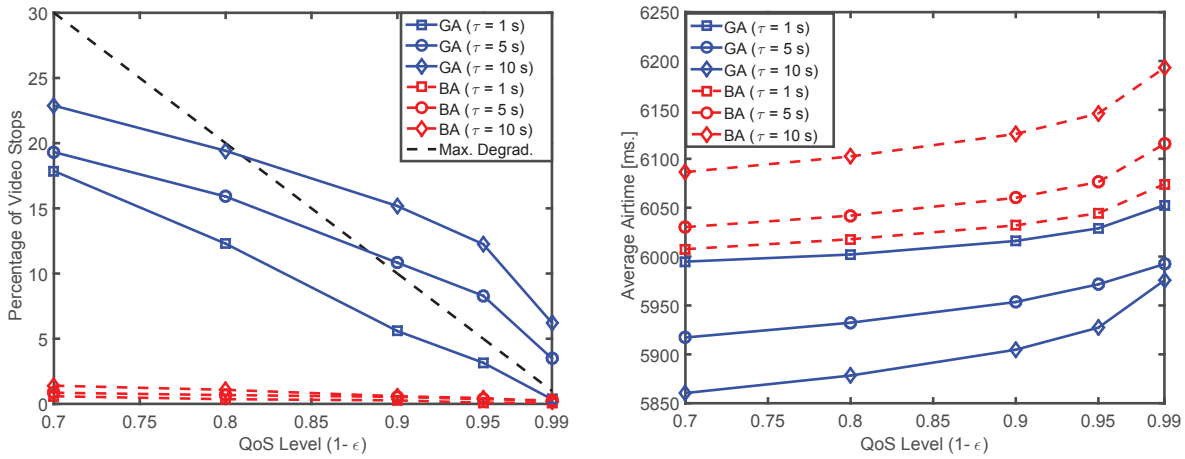
A. Simulation Set-up

We simulate the proposed robust PRA using our modified ns-3 LTE module [26] that is integrated with Gurobi solver [27] to solve the SoCP optimization problem using Barrier method.

Simulation results are averaged over 50 runs with different shadowing values. Two mobility scenarios were considered; urban and rural. Users move at a low speed with small inter-vehicle distances in the urban scenario, and thus experience similar average rate values at the same time interval. The rural scenario models high speed moving vehicles with large inter-vehicle distances. Consequently, users experience different data rates from each other at the same time interval. Video content is then requested by all users at a fixed streaming rate over the considered time horizon. The numerical values of all the parameters are summarized in Table I, while the variance and bounds of each rate are calculated using the previously discussed Monte-Carlo simulation.

B. Evaluation Metrics

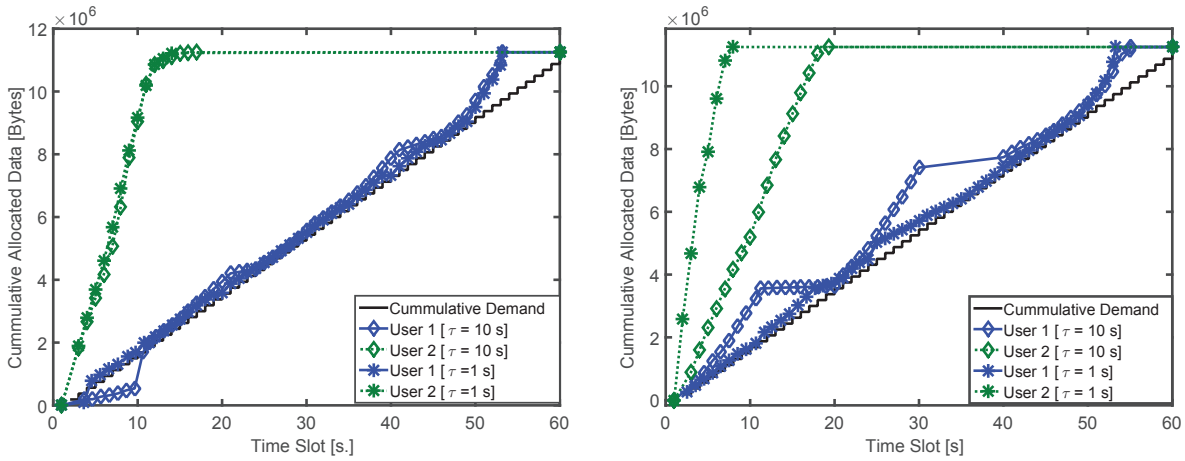
The introduced robust PRA framework is evaluated based on two metrics. The first is the percentage of videos stops



(a) Average Percentage of video stops.

(b) Average BS airtime.

Fig. 4. Performance of the robust framework for varying QoS levels $(1-\epsilon)$ for 2 users experiencing NLoS variance in urban area.



(a) Gaussian approximation

(b) Bernstein approximation

Fig. 5. Allocation at different feedback intervals for 2 users experiencing slow fading with LoS variance. in rural area

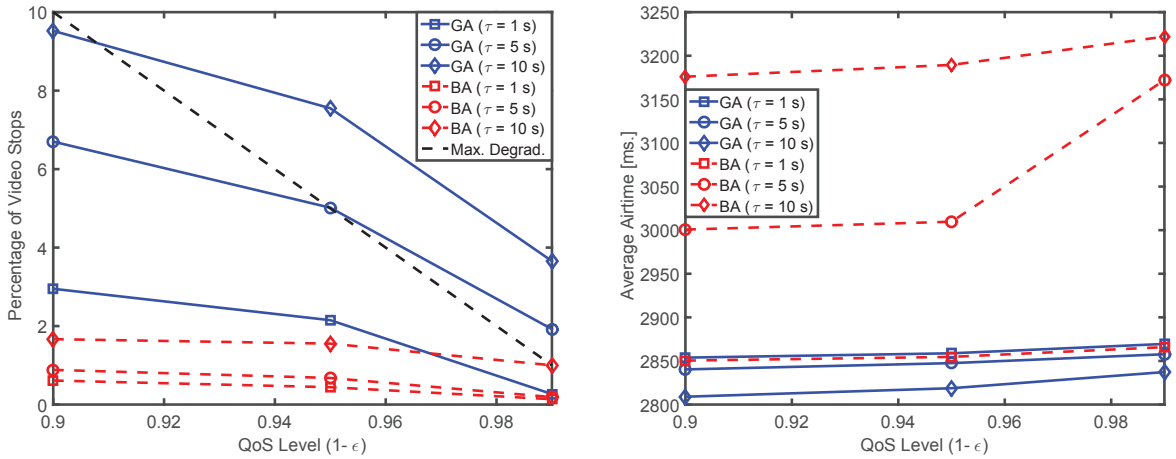
which is the percentage of time slots in which C1 in Eq. 2 is violated [22]. The maximum allowed percentage of stops is equal to $\epsilon \times 100\%$. The second metric is the average BS airtime which reflects the energy consumption in the network [28].

C. Simulation Results

1) Robustness in the Urban Scenario: In urban areas, users start moving from the cell edge towards the centre. In order to decrease computational complexity of the solver, the feedback time τ was set firstly to a relatively long interval equal to 10 s. This is the interval over which the solver recalculates the allocation of all users for the remaining future time slots. In case of GA, the maximum degradation was surpassed for high QoS (i.e., $1-\epsilon \geq 0.9$) as shown in Fig. 4(a). This performance is attributed to the overlooked dependency between the QoS constraints over time. Consequently, demand violation at a certain slot will propagate and affects the satisfaction in the next slots within the feedback interval. Such violations last until reallocation is done for the next slots. The value of τ was set to lower values $\tau = 1$ and $5s$. where less degradation occurs Fig. 4(a), but at the expense of both: increased airtime Fig. 4(b) and the computational complexity.

The BA approach is very conservative, and thus the percentage of stops was kept below the maximum threshold for all the QoS levels and feedback values of τ as shown in Fig. 4(a). However, the airtime performance with τ is inverted compared GA. This is due to the fact that users are moving from a region of low rate towards the cell peak, and BA requires fast feedback to decrease the conservative allocation at the cell edge which consumes large airtime. Large feedback durations continue to allocate large amounts of data at the cell edge.

BA requires small feedback durations to correct its conservative allocation. Similarly, GA also requires the same small feedback time but to recover the degradation in any timeslot and prevent it from affecting the coming ones. The allocation for user 1 in Fig. 5(a) demonstrates the aforementioned properties. In GA Fig. 5(a) where degradation occurs at the



(a) Average Percentage of video stops.

(b) Average BS airtime.

Fig. 6. Performance of the robust framework for varying QoS levels $(1-\epsilon)$ for 2 users experiencing LoS variance in rural area.

first time slot, the small feedback ($\tau = 1\ s$) was able to recover this by recalculating the allocation at the next time slot ($t = 2\ s.$). On the other hand, Bernstein's conservatism avoided the degradation in any of the time slots. However, airtime allocation at early slots (where the rate is minimal) was avoided by frequent feedback, while allocation continues conservatively (large gap above the demand) for the case of $\tau = 10\ s$ as depicted in Fig. 5(b).

2) Robustness in Rural Scenario: The above conclusions were drawn for the case of users experiencing similar radio conditions at the same time. Thus, very conservative solutions only affects the optimality of each user individually. We now consider the rural scenario where some users are located as the cell edge while others are at the cell peak and moving towards the edge. Minimal allocation, to satisfy the QoS, is performed for the users at the cell edge while prebuffering is done for the cell peak users to avoid allocation at future low rate locations. In this scenario, the conservatism of cell edge users is more severe and affects the optimality of cell peak users as well due to the provided small airtime for prebufffering. An example of such a case is shown for user 2 (located at cell peak) in Fig. 5(b). Due to the conservative allocation of user 1 located at cell edge for $\tau = 10s.$, user 2 was unable to prebuffer in the first 10 seconds while located at the cell peak. Thus, the peak user had to wait until reallocation of the cell edge user at $t = 10s.$ so more airtime is provided for the former to prebuffer at relatively lower rates.

Accordingly, the cost of conservatism in the rural scenario has increased and thus the energy gap expanded between Bernstein at ($\tau = 5$ and $10\ s.$) and the less conservative cases: i.e., Bernstein ($\tau = 1\ s.$) and Gaussian as shown in Fig. 6(a). The frequent feedback of Bernstein (i.e. $\tau = 1\ s.$) was able to overcome its expected conservatism and thus results in nearly equal energy consumption compared to the Gaussian case at the same feedback interval. Moreover, the QoS satisfaction of large feedback intervals ($\tau = 5$ and $10\ s.$) is slightly enhanced for the Gaussian case where violation of the maximum degradation occurs only at the highest QoS level for $\tau = 5\ s$, and at the highest two QoS values for $\tau = 10\ s.$ as depicted in Fig. 6(b). This is attributed to the prebuffering strategy for the cell peak users and thus their QoS satisfaction never fails resulting in lower average violation.

TABLE II
PERFORMANCE OF ALL SCHEDULING FOR 4 USERS EXPERIENCING SLOW FADING WITH NLOS VARIANCE. IN URBAN AREA WITH $\epsilon = 0.01$

	PF	PRA (PK)	PRA (NR)	GA (1 s.)	GA (10 s.)	BA (1 s.)	BA (10 s.)
Stops %	0	0	25.5	0.05	2.7	0	0
Airtime	7350	5550	5600	5670	5650	5750	5900

TABLE III
PERFORMANCE OF ALL SCHEDULING FOR 4 USERS EXPERIENCING SLOW FADING WITH LOS VARIANCE. IN RURAL AREA WITH $\epsilon = 0.01$

	PF	PRA (PK)	PRA (NR)	GA (1 s.)	GA (10 s.)	BA (1 s.)	BA (10 s.)
Stops %	0.12	0	7.5	0.5	5.8	0	0
Airtime	3660	2750	2950	2970	2950	3020	3120

3) Comparison with Other Resource Allocators: The introduced GA and BA approximations for the robust PRA framework are now compared against: 1) Proportional fair (PF) as a form of the opportunistic non-PRA, 2) non-robust PRA [11] denoted as PRA (NR) and, 3) the theoretical benchmark PRA that is aware of the exact rate variations denoted as PRA (FK). Results for both urban and rural scenarios are shown in Table II and Table III respectively. Whilst the non-robust PRA resulted in the largest percentage of video stops, both the robust approximations were able to satisfy the QoS at different feedback intervals. This was done without compromising the airtime significantly. Moreover, the robust techniques preserved the energy saving gain in PRA and thus result in less airtime compared to the non-predictive PF. Finally, compared to the benchmark which has full knowledge of the varying rates, small intervals ($\tau = 1\ s.$) provided a comparable airtime saving gain and nearly the same QoS. On the other

hand, higher intervals either compromises the QoS or the energy saving in case of Gaussian and Bernstein respectively due to the reasons previously discussed. We remark that the prediction gain decreases in the second scenario since half the number of users are allocated at the cell peak. Thus, the strategy of both predictive and non-predictive allocators is the same and appears to prebuffer the video of the cell peak users.

VI. Conclusion

We introduced a *robust* PRA framework for energy-efficient video delivery using the Gaussian approximation (GA) and Bernstein approximation (BA) approaches. The scheme was evaluated for different user mobility scenarios and target QoS levels. Simulation results show the resilience of the proposed PRA framework in meeting QoS constraints, while maintaining low energy consumption levels. The BA formulation successfully satisfied the QoS in all movement scenarios and levels $(1-\epsilon)$, regardless the feedback interval τ. However, airtime minimization was suboptimal, particularly for a large τ. Therefore, to minimize energy with BA, a small feedback interval is recommended at the expense of increased computational complexity. On the other hand, the GA formulation provides more energy savings, which are inversely proportional to the feedback intervals. Thus, energy minimization with low complexity can be simultaneously achieved. This comes at the cost of QoS violations especially during higher network loads. In summary, since small feedback intervals should be practically avoided, the BA approach can be applied for high QoS guarantees, while GA is recommended when the energy minimization is the primary objective.

Our future work considers the following enhancements to the robust PRA framework:

1. Decreasing the complexity of low feedback intervals by applying directed heuristic optimization techniques that exploit the problem's features [11].
2. Considering the *joint* probability between the consecutive QoS constraints over the time horizon. This may facilitate both energy minimization and QoS guarantees using the GA with large feedback intervals.
3. Providing less conservative upper bounds for the cumulative generating function to enhance BA's conservatism.

Acknowledgment

This research is supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

[1] L. Correia, D. Zeller, O. Blume, D. Ferling, A. Kangas, I. Godor, G. Auer, and L. Van der Perre, "Challenges and enabling technologies for energy aware mobile radio networks," *IEEE Commun. Magazine*, vol. 48, no. 11, pp. 66–72, 2010.

[2] M. Elazab, A. Noureldin, and H. Hassanein, "Integrated cooperative localization for connected vehicles in urban canyons," in *Proc. IEEE GLOBECOM*, 2015.

[3] A. Mahmoud, A. Noureldin, and H. Hassanein, "Vanets positioning in urban environments : A novel cooperative approach," in *Proc. IEEE VTC*, 2015.

[4] C. Song, Z. Qu, N. Blumm, and A. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, 2010.

[5] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific reports*, vol. 3, no. 2923, pp. 1–9, 2013.

[6] A. Schulman, V. Navda, R. Ramjee, N. Spring, P. Deshpande, C. Grunewald, K. Jain, and V. N. Padmanabhan, "Bartendr: a practical approach to energy-aware cellular data scheduling," in *Proc. ACM Mobicom*, pp. 85–96, 2010.

[7] H. Abou-zeid, H. S. Hassanein, Z. Tanveer, and N. AbuAli, "Evaluating mobile signal and location predictability along public transportation routes," in *Proc. IEEE WCNC*, pp. 1195 – 1200, 2015.

[8] H. Abou-zeid, H. S. Hassanein, and S. Valentin, "Energy-efficient adaptive video transmission: Exploiting rate predictions in wireless networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2013 – 2026, 2014.

[9] H. Abou-zeid and H. S. Hassanein, "Toward green media delivery: location-aware opportunities and approaches," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 38–46, 2014.

[10] Z. Lu and G. de Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *Proc. IEEE INFOCOM*, pp. 2806–2814, 2013.

[11] H. Abou-zeid and H. S. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 92–99, 2013.

[12] H. Abouzeid and H. S. Hassanein, "Efficient lookahead resource allocation for stored video delivery in multi-cell networks," in *Proc. IEEE WCNC*, pp. 1909–1914, 2014.

[13] R. Margolies, A. Sridharan, V. Aggarwal, R. Jana, N. Shankaranarayanan, V. A. Vaishampayan, and G. Zussman, "Exploiting mobility in proportional fair cellular scheduling: Measurements and algorithms," in *Proc. IEEE INFOCOM*, pp. 1339–1347, 2014.

[14] P. Kali and S. W. Wallace, *Stochastic programming*. Springer, 1994.

[15] X. Wang, D. Wang, H. Zhuang, and S. D. Morgera, "Fair energy-efficient resource allocation in wireless sensor networks over fading tdma channels," *IEEE J. Select. Areas Commun.*, vol. 28, no. 7, pp. 1063–1072, 2010.

[16] A. Charnes and W. W. Cooper, "Chance-constrained programming," *Management science*, vol. 6, no. 1, pp. 73–79, 1959.

[17] A.-C. So and Y. J. Zhang, "Distributionally robust slow adaptive ofdma with soft qos via linear programming," *IEEE J. Select. Areas Commun.*, vol. 31, no. 5, pp. 947–958, 2013.

[18] M. J. Abdel-Rahman, F. Lan, and M. Krunz, "Spectrum-efficient stochastic channel assignment for opportunistic networks," in *Proc. IEEE GLOBECOM*, pp. 1272–1277, 2013.

[19] A. Nemirovski and A. Shapiro, "Convex approximations of chance constrained programs," *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 969–996, 2006.

[20] A. Ben-Tal and A. Nemirovski, "Selected topics in robust convex optimization," *Mathematical Programming*, vol. 112, no. 1, pp. 125–158, 2008.

[21] N. Bui and J. Widmer, "Modelling throughput prediction errors as gaussian random walks," in *Poc. IEEE KuVS Workshop*, 2014.

[22] R. Atawia, H. Abou-zeid, H. Hassanein, and A. Noureldin, "Robust resource allocation for predictive video streaming under channel uncertainty," in *Proc. IEEE GLOBECOM*, pp. 4683–4688, Dec 2014.

[23] N. Kolehmainen, J. Puttonen, P. Kela, T. Ristaniemi, T. Henttonen, and M. Moisio, "Channel quality indication reporting schemes for utran long term evolution downlink," in *Proc. IEEE VTC*, pp. 2522–2526, 2008.

[24] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); physical layer procedures," Tech. Specification TS 36.213 v12.5.0, 2015.

[25] 3GPP, "LTE; evolved universal terrestrial radio access (E-UTRA); further advancements for E-UTRA physical layer aspects," Tech. Rep. TR 36.814 V9.0.0, 2010.

[26] H. Abou-zeid, H. S. Hassanein, and R. Atawia, "Towards mobility-aware predictive radio access: modeling; simulation; and evaluation in lte networks," in *Proc. ACM MSWiM*, pp. 109–116, 2014.

[27] Gurobi, "Gurobi Optimization." http://www.gurobi.com/products/features-benefits/. Accessed Mar. 29th, 2015.

[28] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, and H. Holktamp, "Flexible power modeling of lte base stations," in *Proc. IEEE WCNC*, pp. 2858–2862, 2012.