System Level Optimization in Wireless Networks: Managing Interference with Robust Optimization

Sungho Yun, *Student Member, IEEE,* and Constantine Caramanis, *Member, IEEE*

***Abstract*—We consider a robust-optimization-driven system-level approach to interference management in a cellular broadband system operating in an interference-limited and highly dynamic regime. Here, base stations in neighboring cells (partially) coordinate their transmission schedules in an attempt to avoid simultaneous max-power transmission to their mutual cell edge. Limits on communication overhead and use of the backhaul require base station coordination to occur at a slower time scale than the customer arrival process.**

The central challenge is to properly structure coordination decisions at the slow time scale, as these subsequently restrict the actions of each base station until the next coordination period. A further challenge comes from the fact that over longer coordination intervals, the statistics of the arriving customers, e.g., the load, may themselves vary or be only approximately known. Indeed, we show that performance of existing approaches degrades rapidly as the uncertainty in the arrival process increases.

In this paper we show that a two-stage robust optimization framework is a natural way to model two-time-scale decision problems. We provide tractable formulations for the base-station coordination problem, and show that our formulation is robust to fluctuations (uncertainties) in the arriving load. This tolerance to load fluctuation also serves to reduce the need for frequent re-optimization across base stations, thus helping minimize the communication overhead required for system level interference reduction. Our robust optimization formulations are flexible, allowing us to control the conservatism of the solution. Our simulations show that we can build in robustness without significant degradation of nominal performance.

***Index Terms*—Robust Optimization, System Level Optimization, Interference Management.**

I. INTRODUCTION

In down-link cellular systems with small cells and full frequency re-use, base station service rates are coupled by inter-cell interference, thus jeopardizing their performance. Thus minimizing the effect of inter-cell interference is paramount. Rather than physical-level approaches that require millisecond-scale inter-base-station coordination, i.e., at the time scale of channel fluctuations, (e.g., zero-forcing) or frequency partitioning (increasingly undesirable as the number of base stations increases), we consider here a *system level* algorithmic approach, that aims to determine an optimal scheduling policy of the whole system simultaneously, to minimize joint transmissions of base stations to their mutual boundary. While avoiding the requirement to communicate at the millisecond time scale, a naive implementation of such a scheme would require full knowledge of the behavior of arriving customers over the entire system at all time, thus requiring base stations to communicate all information of any change in the spatial load – arguably prohibitively high inter-base station communication overhead. To overcome this difficulty, Rengarajan and de Veciana [1] propose a novel framework in which customers are aggregated into classes, and then base stations coordinate at this coarse level, jointly optimizing a transmission schedule using only the statistics of customer (class) arrivals, and offered load, allowing base-station coordination to occur at a much slower time scale than customer arrival.

When the offered load is known at the time of coordination, the techniques proposed in [1] have been shown to increase the stability region, decrease file transfer delay significantly at all load levels, while also increasing the uniformity in the coverage. As we demonstrate in the sequel, this exact knowledge of the offered load at each time seems to be crucial, as performance quickly degrades as uncertainty in the offered load increases.

In this paper we develop tools from multi-stage robust and stochastic optimization to tackle this problem of load uncertainty. Most approaches to solve resource allocation problems in wireless networks are based on exact knowledge of environment variables, e.g., channel state information, offered load, noise power, etc. as in [2], [3] and [4] and some use robust optimization paradigms to handle uncertain random variables, see e.g. [5]. However, to the best of our knowledge, there have been no attempts to model adaptability by using techniques from multi-stage adjustable robust and stochastic optimization in wireless cellular systems.[1]

Our first contribution is in showing that multi-stage (for our purposes here, two-stage) robust and stochastic optimization provide the right optimization framework for considering distributed decision-making with coordination and uncertainty at different time scales in wireless cellular systems. We consider uncertainty in the load, as well as uncertainty in its distribution among base stations. We formulate tractable (in particular, convex) optimization formulations for robust coordination considering both capacity-maximizing and delay-minimizing formulations. In extensive simulation experiments, we show that our robust and stochastic optimization formulations successfully immunize our solutions to variations in the load, both in terms of the stability region, and also in terms of average delay. At the same time, we investigate the

S. Yun and C. Caramanis are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, 78712 USA e-mail: {shyun,caramanis}@mail.utexas.edu.

This work was supported in part by NSF grants and DTRA grants.

[1]There has been some work in network provisioning, although quite different from what we consider here, e.g., see [6] where a two-stage robust optimization approach is used to solve a network flow problem.

so-called price of robustness, or conservatism, by considering the degradation of the nominal performance (how well does the robust solution do when there are no variations?). Our simulation results indicate that we can build in robustness without significant degradation from the nominal performance (i.e., when there is no deviation, the robust solution does not sacrifice much in terms of performance).

We focus on two models for the variation in the customer arrival process. In the first, we assume that the arrival rate to base station cells is fixed, while the distribution of customers and their location within each cell varies. This models the scenario where total traffic is unaffected, but some event leads to variation in traffic distribution patterns. Next, we consider the case where the distribution of customers among base station cells, and within cells is as expected, but the aggregate level of traffic varies. This models a scenario where customer location is unaffected, but some event leads to increased or decreased total traffic.

The structure of this paper is as follows. In Section II we provide the basic system model and setup of our problem. In Section III, we briefly summarize some previous results and give background on Adjustable Robust Optimization and Stochastic Optimization problems. In Section IV-A, we motivate the robust and stochastic optimization counterpart of the system-level network optimization problem [1]. In Sections IV-B, IV-C and IV-D, we consider three different formulations, each designed for different models for the uncertainty of the arrival process, and different objective functions, namely, capacity maximization, and delay minimization. We provide simulations to illustrate the performance of each. Finally, Section V concludes this paper.

II. System Model and Setup

Rengarajan and de Veciana [1] propose a novel solution to enhance wireless broadband capacity in the case where neighboring cell transmissions are the limiting source of interference. To take advantage of the diversity in users' sensitivity to interference from the neighboring cells, they first group customers in each base station into several classes according to their interference sensitivities and system loads. Then base stations jointly optimize a transmission policy determining which class of customers each base station should serve at each time and at what power level, *using only statistics of channel quality and load distribution*. Their model is the starting place for our work here, and therefore we use similar notation to theirs wherever possible.

Let N be the number of base stations. Each base station, b, serves customers that are aggregated into K_b customer classes. We assume that each customer is served by a single base station. The clustering of customers into a smaller number of classes serves to reduce the complexity of the required communication and coordination, as well as the computational complexity of the resulting optimization problems. We adopt the same aggregation model and approach as in [1], and hence refer the reader there for the details. Arrivals to class k of base station b are Poisson with rate λ_{bk} and mean file size $\overline{F}_{bk}$. Thus the offered load is $\rho_{bk} = \lambda_{bk}\overline{F}_{bk}$.

The operating decisions at each base station consist of the transmit power level, and the customer class served. The interference seen at a particular cell, depends only on the power level at which its neighboring base stations are transmitting, and in particular does not depend on the customer class being served.

This simple observation forms the basis for the low-overhead coordination schemes we propose: base stations use customer load statistics to coordinate via slow-scale usage of the backhaul, deciding on a power schedule (who will transmit at what power, at what time). This decouples the the actions of neighboring base stations, allowing each individual base station to serve its customer classes optimally, subject to the agreed-upon power constraint, in order to best serve the realization of customer load at each moment. Crucially, no information about the realized customer load need be communicated to neighboring base stations, eliminating the need for high-bandwidth, low-lattency backhaul usage.

This paper is about optimally deciding upon the slow-scale power-level schedule, when the offered load $\vec{\rho} = \{\rho_{bk}\}$ is not known exactly. We define "optimality" with respect to two objectives: maximizing the stability of the system, i.e., the offered load at which delay remains bounded, and also minimizing average customer delay at all loads. We approach these two tasks by formulating optimization problems modeling capacity maximization and delay minimization. While delay minimization is directly related to the quantity we wish to control, capacity maximization is an indirect proxy that offers computational advantages because it can be expressed as a linear function of our decision variables. This is particularly true for our two-stage optimization formulations.

We now give the single stage optimization formulation with no uncertainty, using a generic objective function, to set the stage for the multi-stage formulations in the following sections.

We assume that there are L total joint power profiles at which the base stations can transmit, and we define decision variables $\vec{\alpha} = \{\alpha_l\}_{l=1}^{L}$ to denote the fraction of time the base stations spend broadcasting at each power profile. The reason we require joint power profile variables, rather than individual power profile variables for each base station, is that this will enable power profile coordination across base stations. We define decision variables $\vec{p} = \{p_{bk}^{l}\}$ to denote the fraction of time base station b serves customer class k, when joint power profile $l \in \{1, \ldots, L\}$ is being used.

As defined above, let $\vec{\lambda}$ denote the customer arrival rate, so that $\vec{\rho}$ is the vector of offered loads. In the formulation below, there is no uncertainty in $\vec{\lambda}$ (and hence in the offered load). Let $f(\vec{\alpha}, \vec{p})$ denote the objective function. The objective function $f(\vec{\alpha}, \vec{p})$ is at this point generic, but below we provide two formulations, differing only in our choice of f, with one modeling capacity maximization, the other delay minimization.

We obtain the following optimization problem:

$$\begin{array}{lll} \min_{\vec{\alpha},\vec{p}} & f(\vec{\alpha},\vec{p}) & \\ \text{s.t.} & \rho_{bk} \le R_{bk}(\vec{\alpha},\vec{p}) & \forall b,k \\ & \sum_{k=1}^{K_b} p_{bk}^{l} \le \alpha_l & \forall b,l \\ & \sum_{l=1}^{L} \alpha_l \le 1 & \\ & p_{bk}^{l} \ge 0 & \forall b,k,l \\ & \alpha_l \ge 0 & \forall l \end{array} . \quad (1)$$

We use $R_{bk}(\vec{\alpha},\vec{p})$ to denote the capacity allocated to class k in base station b by schedule $\vec{\alpha},\vec{p}$. Therefore the constraints $\rho_{bk} \le R_{bk}(\vec{\alpha},\vec{p})$ are the system stability constraints. The remaining constraints merely say that the fractions of time spent in each power profile cannot add to more than 100% of time, and that amount of time each base station wishes to serve customers under power profile l must not exceed the total amount of time allocated to power profile l, namely, α_l.

In this formulation, the schedule of power profiles (or power levels) is determined simultaneously with the decisions of which class each base station will serve. The novel optimization concept we propose here, is to separate power-level decisions from serving-class decisions, into a two-stage design: the former decisions (power profile) are our coordination decisions, taken before the uncertainty is realized, while the latter decisions (serving class) are made only after the uncertainty is realized (i.e., after the base stations see their own realized load in each customer class). These form our first and second stage decisions, as discussed below.

III. Adjustable Robust/Stochastic Optimization

Stochastic and Robust Optimization have recently found successful application in select signal processing and communication problems ([**?**], [**?**]) where they have been important in dealing with parameter uncertainty. One of the main optimization-based contributions of this paper is to introduce multi-stage robust and stochastic optimization as a useful and important modeling methodology and a computational tool for problems in wireless networks. In particular, we aim to illustrate its use in dealing with dynamic problems affected with uncertainty, yet where coordination among agents (in our case, neighboring base stations) occurs at a slower time scale than the realization of the uncertainty (in our case, the variation in customer arrivals).

The purpose of this section is to provide a brief review of robust and stochastic optimization, and then to introduce multi-stage (particularly two-stage) robust and stochastic optimization problems. We first review the single-stage (non-adaptable) robust optimization paradigm. To make the connection to our subsequent discussion on two-stage optimization clear, we write the optimization variables as u and v. In the sequel, the former will denote first-stage decisions to be implemented immediately, while the latter will denote second-stage decisions, implemented at a later time. In our context, first stage decisions will represent coordination decisions made prior to the realization of the customer arrivals (i.e., the joint power level schedule, $\{\alpha_l\}$), and our second stage decisions will represent scheduling decisions made by each base station only after the arrival of customers is realized (which customer classes to serve under each power profile, $\{p_{bk}^{l}\}$).

A. Robust and Stochastic Optimization

Consider a, for now linear, optimization problem:

$$\begin{array}{ll} \min_{u,v} & cu + qv \\ \text{s.t.} & Uu + Vv \le b \\ & u \ge 0 \\ & v \ge 0. \end{array} \quad (2)$$

Robust and Stochastic Optimization address the setting wherein the parameters U and V are only partially known, or are noisy or corrupted.[2] Stochastic optimization treats the uncertainty as driven by a distribution, and requires the constraints to be satisfied in some probabilistic sense, either with high probability, or by penalizing constraint violation and then minimizing the expected penalty (e.g., [**?**], [**?**], [11]). Robust optimization, on the other hand, assumes a *deterministic, set-based* uncertainty model.[3] Using ω to denote uncertainty, the robust optimization formulation of the nominal problem above takes the following form:

$$\begin{array}{ll} \min_{u,v} & cu + qv \\ \text{s.t.} & U(\omega)u + V(\omega)v \le b \quad \forall \omega \in \Omega \\ & u \ge 0 \\ & v \ge 0 \end{array} \quad (3)$$

For a variety of uncertainty sets, Ω, the above problem is efficiently solvable and in fact is again a linear program for polyhedral Ω. For further details, we refer the reader to [9], [10] and the survey [**?**] and references therein.

Whether or not ω is treated as a stochastic or a deterministic (worst-case) parameter, a key element of the formulation is that decisions u and v are fixed *before* the realization of the uncertainty, ω. If, however, the second stage decisions, v, are implemented only *after* the realization of the uncertainty, ω, i.e., if the decision maker can observe ω before implementing v, then the above formulation is potentially very conservative. It means that the decision-maker is giving up any opportunity to adapt to the realized uncertainty. One may also consider the so-called *receding-horizon* approach, wherein one solves the above problem, *and then reoptimizes* once additional information about the uncertainty becomes known. This approach has two disadvantages: first, it requires potentially significantly increased computation since optimization problems must be solved at the faster time scale, and second, the first-stage decisions may still be suboptimal, since they are obtained by solving the above problem that does not explicitly take into account any second-stage adaptability.

B. Two-Stage Optimization

The two-stage optimization paradigm addresses precisely these two short-comings of the above model. It allows us to decide only the first stage, non-adjustable variables u, and

[2]The case of unknown c, q, or b can be treated similarly, through a simple transformation of the problem.

[3]Numerous papers have shown that despite the worst-case formulation of Robust optimization as opposed to the probabilistic nature of the uncertainty in stochastic optimization, the performance of solutions to the former typically compares very favorably to that of the latter. Indeed, the two formulations are often differentiated by suitability of the formulation, and as we discuss at length in this paper, tractability of the resulting optimization problem.

instead of optimizing over v, we can optimize over *policies for adaptability* for the second stage solution. Moving the objective function into the constraints, we can write this fully-adaptable optimization formulation as:

$$\begin{array}{ll} \min_{u,\gamma} & \gamma \\ \text{s.t.} & \exists\, u \text{ such that } \forall \omega \in \Omega, \\ & \exists\, v \text{ with } \left\{ \begin{array}{l} \gamma \geq cu + qv(\omega) \\ U(\omega)u + V(\omega)v(\omega) \leq b \\ u \geq 0 \\ v(\omega) \geq 0 \end{array} \right\} \end{array} \tag{4}$$

Equivalently, this can be written as:

$$\begin{array}{lll} \min_u \quad \max_\omega & cu + qv(\omega) & \\ \quad \text{s.t.} & U(\omega)u + V(\omega)v(\omega) \leq b & \forall \omega \\ & u \geq 0 & \\ & v(\omega) \geq 0 & \forall \omega \end{array} \tag{5}$$

The second formulation is an optimization over first stage decisions, and second stage functions of uncertainty, or policies. Following [7], [8], we refer to these equivalent formulations as the *Adaptable Robust Counterpart* (ARC).

Similarly, we may consider the two-stage model for stochastic optimization, where the uncertainty, ω, is chosen according to some distribution. In our setting, constraints correspond to stability of our system, and therefore we must impose these constraints deterministically, obtaining a cross between the robust and stochastic optimization viewpoints. Thus, the resulting multi-stage optimization problem takes the form:

$$\begin{array}{lll} \min_u & cu + E_\omega[\min_{v(\omega)} qv(\omega)] & \\ \text{s.t.} & U(\omega)u + Vv(\omega) \leq b & \forall \omega \\ & u \geq 0 & \\ & v(\omega) \geq 0 & \forall \omega. \end{array} \tag{6}$$

In general, for both robust and stochastic optimization formulations, obtaining the optimal policy $v(\omega)$ exactly is intractable, and we must be content to look for solutions in restricted classes of functions. In this paper, we consider two techniques for accomplishing this in a computationally tractable manner. First, following [7] and adapting techniques that have been used in inventory management problems, we consider affine functions of the uncertainty, ω:

$$v(\omega) = w + W\omega.$$

While even this restriction may in general be NP-hard, we show that in our application, such an affine rule is (a) tractable, reducing to a linear optimization, (b) easily implemented at the fast time-scale, requiring only evaluation of an affine function, thus much more computationally efficient than reoptimization required in the receding horizon approach, and (c) in one important case we consider (see Section IV-B), this affine policy does not restrict the adaptability and hence for this case is equivalent to the ARC.

We next consider an approach allowing for more general non-affine functions of the uncertainty. Here, we essentially fit a piecewise bilinear function to the optimal adaptability function. This approach is important for the case where affine adaptability is not optimal, or the resulting optimization problem for affine adaptability is not tractable. We show that our piecewise bilinear approach can be computed via a tractable, convex optimization problem, and further show that its performance is very close to that of optimal adaptability. Like affine adaptability, this approach also has the implementation advantage of not requiring reoptimization at the second stage, and hence at the fast time scale, thus providing significant computational advantages over the receding horizon approach.

C. Advantages of Different Two-Stage Formulations

We have discussed several variants of stochastic and robust optimization, with different formulations of adaptability. Also, as mentioned in the introduction, we consider two different noise models for the customer arrival process. Some combinations of uncertainty model and adaptability formulation prove more tractable than others. We try to elucidate this in the following section. In addition, we try to illustrate the breadth of the modeling and computational techniques we introduce, even though space constraints make it impossible to exhaustively explore every combination of adaptability formulation and uncertainty model.

IV. Uncertain Arrival Rates

When the loads are known exactly, simulations reveal that the solution of [1] demonstrates remarkable improvements over a simple baseline no-coordination solution. Yet these gains deteriorate when the offered loads change over time at a faster scale than the base stations can re-optimize. In this paper, by using robust and stochastic optimization techniques ([7], [9], [10] and [11]), we propose an approach to make the solution robust to the changes of the offered loads.

A. Two-stage Optimization

In [1], base stations coordinate, choosing joint power-and-class transmission schedules. We consider separately the two different elements of the transmission profile: a power profile and a class profile. The power profile represents the transmit power level for each base station; the class profile represents the class that each base station will serve. Since the interference level seen by a base station depends only on the power level of its neighbors, and not which classes they might be serving, fixing a power profile also fixes interference to each base station. Therefore class scheduling decisions become decoupled as soon as power profiles are fixed, and hence inter-base-station communication is not needed beyond power profile scheduling. Hence, the two-stage optimization setting becomes natural: before the actual offered loads for each class in each base station become known, base stations coordinate and decide upon the power profile schedule; next base stations decide on the class profile schedule after the offered loads become known without further communication with other base stations.

This two-stage formulation allows us to consider robustness to uncertainty in the offered load. We consider two different uncertainty models for the variation in the offered loads, exploiting the strengths of robust and stochastic optimization, respectively. The first model is a *fixed total arrival rate*

model: while the actual arrival rates of user classes fluctuate in each base station, the sum of fluctuations is always zero. We handle this using the robust optimization paradigm: the rates fluctuate arbitrarily *in a given uncertainty set*, and the solution must be robust to all allowed variations. The second model is a *fixed arrival rates ratio* model in which the total arrival rate fluctuates while the fraction of each arrival rate for each base station and class is fixed. We use a stochastic optimization approach. The arrival rate varies according to a stochastic process, and the solution minimizes the expected customer delay. In both models, we deterministically enforce stability constraints. We believe that many other models, and combinations of the ones we treat here, can be approached using the methods we present in this paper.

We illustrate the main two-stage optimization formulation using the robust model. The stochastic model is considered in detail in Section IV-C. Let $\vec{\lambda}$ denote the (unknown) offered load, varying in an uncertainty set $\mathcal{Z}$. The decision variables $\{\alpha_l\}_{l=1}^L$ represent the joint decisions on power profile coordination, with each $l = 1, \ldots, L$ denoting a different joint power profile. The second stage decisions are given by $p_{bk}^l(\vec{\lambda})$. Note that unlike the formulation in (1), here these decisions depend on the realization of the uncertainty, $\vec{\lambda}$. This explicit dependence on $\vec{\lambda}$ indicates that they are second-stage decisions, made after the uncertainty realization. The variables $\{\alpha_l\}_{l=1}^L$, on the other hand, have no dependence on $\vec{\lambda}$, as they are made in the coordination phase, before the realization of the uncertainty. We write the variables $p_{bk}^l(\vec{\lambda})$ as general functions here for clarity of exposition. To solve the optimization, we must restrict the class of functions, so as to maintain tractability (in particular, convexity) of the problem. In the next two sections (for both stochastic and robust formulations), we restrict to *affine functions of the uncertainty*. We obtain the following robust optimization problem, which is the robust analog to the optimization given in (1) above:

$$\begin{array}{lllll} \min_{\vec{\alpha},\vec{p}(\vec{\lambda})} & f(\vec{\alpha},\vec{p}(\vec{\lambda})) & & & \\ \text{s.t.} & \rho_{bk} \le R_{bk}(\vec{\alpha},\vec{p}(\vec{\lambda})) & \forall b,k & \forall \vec{\lambda} \in \mathcal{Z} & \\ & \sum_{k=1}^{K_b} p_{bk}^l(\vec{\lambda}) \le \alpha_l & \forall b,l & \forall \vec{\lambda} \in \mathcal{Z} & \\ & \sum_{l=1}^{L} \alpha_l \le 1 & & & \\ & p_{bk}^l(\vec{\lambda}) \ge 0 & \forall b,k,l & \forall \vec{\lambda} \in \mathcal{Z} & \\ & \alpha_l \ge 0 & \forall l. & & \end{array} \quad , \quad (7)$$

As in (1), $R_{bk}(\vec{\alpha},\vec{p}(\vec{\lambda}))$ denotes the capacity allocated to class k in base station b by schedule $\vec{\alpha},\vec{p}(\vec{\lambda})$[4]. System stability is enforced in the first constraint. The second constraint enforces consistency of the class scheduling decisions with respect to the first stage power profile schedule. The third constraint says that the power profile schedule cannot take more than 100% of time, and the final two nonnegativity constraints say that fractions of time must be nonnegative.

Note that with a singleton uncertainty set, we recover the original single-stage formulation in (1), and the original formulation in [1].

[4]If base stations use processor sharing to serve users within each class, R_{bk} is given by a harmonic mean (see [1]). Optimizing using the harmonic mean may be difficult, as it is nonconvex, and for this reason, we use an arithmetic mean approximation.

As mentioned above, we use two different objectives for the optimization problems following [1]. Minimizing $f = \sum_{l=1}^L \alpha_l$ corresponds to a capacity maximizing schedule, while $f = \sum_{b=1}^N \sum_{k=1}^{K_b} \frac{\frac{\rho_{bk}}{R_{bk}(\vec{\alpha},\vec{p}(\vec{\lambda}))}}{1-\frac{\rho_{bk}}{R_{bk}(\vec{\alpha},\vec{p}(\vec{\lambda}))}}$ minimizes the total average delay. Because capacity maximization can be expressed as a linear function of the decision variables, it offers computational advantages. However, the resulting power profile schedule and class schedule has worse average-delay performance than the schedule resulting from minimizing the delay explicitly.

We use the formulation above, and both the capacity maximization and delay minimization objective functions, combined with different optimization paradigms in different uncertainty models, to explore tractability, effectiveness and applicability of various approaches to our problem. As space limitations prohibit an exhaustive exploration of all combinations of optimization model, noise model, and objective function, we state why we choose one over the other for each models:

1) First, in Section IV-B we consider the case of fixed total arrival rate. We use the robust optimization paradigm, modeling the uncertainty in a deterministic way. We consider capacity maximization, and show that in this case the optimal adaptable functions are in fact affine. Thus we obtain the optimal adaptable functions solving a convex optimization problem. This is largely made possible because capacity maximization is a linear objective function.
2) Next, in Section IV-C, we consider the case of uncertain total arrival rate, but fixed ratio across customer classes. Due to the very low dimensionality (1-dimension) of the uncertainty set of the problem which is not the case in 1), we can sample and approximate the distribution of the uncertainty to use a stochastic optimization formulation, minimizing the expected customer delay that results in better average delay than the capacity maximization.
3) Finally, in Section IV-D, we revisit the case of fixed total arrival rate, this time minimizing delay rather than maximizing capacity. Because of the form of the delay minimization objective, a robust optimization formulation for the noise, along with an affine model for adaptability, is no longer tractable. We use a stochastic model for the uncertainty, and develop a new piecewise linear model for the adaptability.

B. Uncertain Arrival Rates Ratio Model with Fixed Total Arrival Rate

1) Assumptions and an Adjustable Counterpart: In this section, we consider a *fixed total arrival rate* model. If load fluctuations are independent across classes, by LLN results, fixed total arrival rate holds in the limit of many classes. While we do not address it in this paper, we note that one can treat the case where the sum of fluctuations is small but not necessarily zero, in a precisely analogous manner.

Let us define an uncertainty set, $\mathcal{Z}_b$ for base station b

$$\mathcal{Z}_b = \left\{ \begin{array}{l} \vec{\lambda}_b = (\lambda_{b1}, \cdots, \lambda_{bK_b}) | \\ \qquad \forall k, \lambda_{bk} \in [(1-\theta)\lambda^*_{bk}, (1+\theta)\lambda^*_{bk}], \\ \qquad \sum_{k=1}^{K_b} \lambda_{bk} = \sum_{k=1}^{K_b} \lambda^*_{bk} \equiv \lambda^*_b \end{array} \right\} \tag{8}$$

where λ^*_{bk} is the nominal arrival rate of user class k in base station b. Thus the true rates must satisfy two properties: individually they cannot deviate by more than $\theta\%$ of the nominal value; and moreover the aggregate deviation must be rate neutral, that is, the sum of the realizations must equal the sum of the nominal rates.

Note that the level of deviation is controlled by the parameter θ. This parameter is under the control of the system designer, who adjusts this θ parameter in order to balance the conservatism and robustness of the solution. Since we enforce feasibility for all realizations in the uncertainty set, larger θ results in a more robust, but also more conservative solution. For $\theta = 0$, we recover the nominal optimal solution, which corresponds to knowing the exact rates to be $\{\lambda^*_{bk}\}$. For each base station b, $\vec{p}$ depends on the arrival rates in its cell. Thus second stage decision variables tune themselves according to the offered loads, and no additional communication overhead is required.

2) Affinely Adjustable Robust Counterpart (AARC): Adaptability allows the second stage solutions to respond to the realized uncertainty. Yet as discussed in Section III-B, for computational reasons, we must restrict the structure of the functions representing the second stage decisions in (7), in order to be able to solve the resulting optimization problem. Restricting $p^{\,l}_{bk}$ to be an affine function of $\vec{\lambda}_b$, as in [7], [12], we have $p^{\,l}_{bk}(\vec{\lambda}_b) = \pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \pi^{\,l}_{bkm}\lambda_{bm}$. Using this formulation, and setting capacity maximization as our objective function, we obtain a linear two-stage robust optimization problem:

$$\begin{array}{lll} \min_{\vec{\alpha},\vec{\pi}} & \sum_{l=1}^{L} \alpha_l & \\ \text{s.t.} & \rho_{bk} \le \sum_{l=1}^{L} (\pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \pi^{\,l}_{bkm}\lambda_{bm}) E_I[R^l_I|b,k] & \\ & \qquad \forall b,k \ , \forall \vec{\lambda}_b \in \mathcal{Z}_b & \\ & \sum_{k=1}^{K_b} (\pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \pi^{\,l}_{bkm}\lambda_{bm}) \le \alpha_l & \\ & \qquad \forall b,l \ , \forall \vec{\lambda}_b \in \mathcal{Z}_b & \\ & \sum_{l=1}^{L} \alpha_l \le 1 & \\ & \pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \pi^{\,l}_{bkm}\lambda_{bm} \ge 0 & \\ & \qquad \forall b,k,l \ , \forall \vec{\lambda}_b \in \mathcal{Z}_b & \\ & \alpha_l \ge 0 & \forall l \end{array} \tag{9}$$

Theorem 1: The AARC (9) is equivalent to the following linear optimization problem.

$$\begin{array}{lll} \min & F & \\ \vec{\alpha},\vec{\pi},\vec{\beta},\vec{\zeta},\vec{\mu},\vec{\sigma},\vec{\nu} & & \\ \text{s.t.} & \beta^{\,l}_{bk0} - \beta^{\,l}_{bkm} \le \pi^{\,l}_{bkm} \le \beta^{\,l}_{bk0} + \beta^{\,l}_{bkm} & \forall b,k,l,m \\ & \zeta_{bkm} = \sum_{l=1}^{L} \pi^{\,l}_{bkm} E[R^l_I|b,k] & \forall b,k,m,k \ne m \\ & \zeta_{bkm} = \sum_{l=1}^{L} \pi^{\,l}_{bkm} E[R^l_I|b,k] - \overline{F}_{bk} & \\ & & \forall b,k,m,k = m \\ & \mu_{bk0} - \mu_{bkm} \le \zeta_{bkm} \le \mu_{bk0} + \mu_{bkm} & \forall b,k,m \\ & \sigma^{\,l}_{bm} = \sum_{k=1}^{K_b} \pi^{\,l}_{bkm} & \forall b,l,m \\ & \nu^{\,l}_{b0} - \nu^{\,l}_{bm} \le \sigma^{\,l}_{bm} \le \nu^{\,l}_{b0} + \nu^{\,l}_{bm} & \forall b,l,m \\ & \sum_{l=1}^{L} \pi^{\,l}_{bk0} E_I[R^l_I|b,k] & \\ & \quad + \sum_{m=1}^{K_b} \zeta_{bkm}\lambda^*_{bm} & \\ & \quad - \theta \sum_{m=1}^{K_b} \mu_{bkm}\lambda^*_{bm} \ge 0 & \forall b,k \\ & \sum_{k=1}^{K_b} \pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \sigma^{\,l}_{bm}\lambda^*_{bm} & \\ & \quad + \theta \sum_{m=1}^{K_b} \nu^{\,l}_{bm}\lambda^*_{bm} \le \alpha_l & \forall b,l \\ & \pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \pi^{\,l}_{bkm}\lambda^*_{bm} & \\ & \quad - \theta \sum_{m=1}^{K_b} \beta^{\,l}_{bkm}\lambda^*_{bm} \ge 0 & \forall b,k,l \\ & \beta^l_{bkm} \ge 0 & \forall b,k,m,l,m \ne 0 \\ & \mu_{bkm} \ge 0 & \forall b,k,m,m \ne 0 \\ & \nu^l_{bm} \ge 0 & \forall b,m,l,m \ne 0 \\ & \sum_{l=1}^{L} \alpha_l \le F & \\ & \sum_{l=1}^{L} \alpha_l \le 1 & \\ & \alpha_l \ge 0 & \forall l \end{array} \tag{10}$$

Proof: Consider the 4^{th} constraint of (9) :

$$\pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \pi^{\,l}_{bkm}\lambda_{bm} \ge 0 \quad \forall b,k,l, \quad \forall \vec{\lambda}_b \in \mathcal{Z}_b.$$

This constraint holds if and only if the optimal value in the following problem

$$\begin{array}{lll} \min_{\vec{\lambda}_b} & \pi^{\,l}_{bk0} + \sum_{m=1}^{K_b} \pi^{\,l}_{bkm}\lambda_{bm} & \\ \text{s.t.} & \lambda_{bm} \ge (1-\theta)\lambda^*_{bm} & \forall m \in [1,\cdots,K_b] \\ & \lambda_{bm} \le (1+\theta)\lambda^*_{bm} & \forall m \in [1,\cdots,K_b] \\ & \sum_{m=1}^{K_b} \lambda_{bm} = \lambda^*_b & \end{array} \tag{11}$$

is nonnegative.

By strong duality for linear programming (e.g., [**?**]), the optimal value is nonnegative if and only if the corresponding dual problem

$$\begin{array}{lll} \max_{\vec{\gamma},\vec{\delta},\xi} & \sum_{m=1}^{K_b} (1-\theta)\lambda^*_{bm}\gamma_m & \\ & \quad - \sum_{m=1}^{K_b} (1+\theta)\lambda^*_{bm}\delta_m + \lambda^*_b\xi + \pi^{\,l}_{bko} & \\ \text{s.t.} & \gamma_m - \delta_m + \xi = \pi^{\,l}_{bkm} & \forall m \in [1,\cdots,K_b] \\ & \vec{\gamma} \ge 0 & \\ & \vec{\delta} \ge 0 & \end{array} \tag{12}$$

has a nonnegative optimal value., i.e., $\exists \vec{\gamma}, \vec{\delta}$, and ξ s.t.

$$\begin{array}{ll} \sum_{m=1}^{K_b} (1-\theta)\lambda^*_{bm}\gamma_m & \\ \quad - \sum_{m=1}^{K_b} (1+\theta)\lambda^*_{bm}\delta_m + \lambda^*_b\xi + \pi^{\,l}_{bko} \ge 0 & \\ \gamma_m - \delta_m + \xi = \pi^{\,l}_{bkm} & \forall m \in [1,\cdots,K_b] \\ \vec{\gamma} \ge 0 & \\ \vec{\delta} \ge 0 & \end{array} \tag{13}$$

Let $\beta^l_{bkm} = \gamma_m + \delta_m$ and $\beta^l_{bk0} = \xi$. Then (13) is equivalent to

$$\begin{array}{l} \sum_{m=1}^{K_b} \lambda^*_{bm}(\pi^l_{bkm} - \beta^l_{bk0}) \\ \quad - \theta \sum_{m=1}^{K_b} \lambda^*_{bm}\beta^l_{bkm} + \lambda^*_b\beta^l_{bk0} + \pi^l_{bko} \geq 0 \\ 2\gamma_m = \beta^l_{bkm} - \beta^l_{bk0} + \pi^l_{bkm} \geq 0 \\ 2\delta_m = \beta^l_{bkm} - \beta^l_{bk0} - \pi^l_{bkm} \geq 0 \end{array} \tag{14}$$

which is equivalent to

$$\begin{array}{l} \pi^l_{bk0} + \sum_{m=1}^{K_b} \pi^l_{bkm}\lambda^*_{bm} - \theta \sum_{m=1}^{K_b} \beta^l_{bkm}\lambda^*_{bm} \geq 0 \\ \beta^l_{bk0} - \beta^l_{bkm} \leq \pi^l_{bkm} \leq \beta^l_{bk0} - \beta^l_{bkm} \\ \beta^l_{bkm} \geq 0 \end{array} \tag{15}$$

Similar arguments can be applied for the rest of the constraints in (9). ∎

Solving the resulting LP, we obtain an affine policy for determining the class profile schedule for each base station, as a function of its local offered load variation.

Moreover we claim that with this affine policy, this Affinely Adjustable Robust Counterpart is actually equivalent to the Adjustable Robust Counterpart of the nominal optimization problem which means that we get no penalty by restricting the second stage policy to be an affine function.

Proposition 1: In this model, the AARC is equivalent to the ARC, and is hence optimal.

Proof: Note that the AARC and the ARC have the same objectives. Therefore it suffices to show the equivalence of constraint sets of those two problems.

Let $\mathcal{X}$(ARC) be the constraint set of ARC and $\mathcal{X}$(AARC) be the constraint set of AARC. First we show that $\mathcal{X}$(ARC)$\supseteq$ $\mathcal{X}$(AARC). Since the ARC has no restriction on the class of second stage policies while the AARC restrict the class of functions into affine functions, obviously, the ARC has a bigger feasible set which includes that of the AARC. Hence, if a solution is feasible to the constraint set of the AARC, it is also feasible to the constraint set of the ARC. Therefore $\mathcal{X}$(ARC)$\supseteq$ $\mathcal{X}$(AARC).

Now we show that $\mathcal{X}$(ARC)$\subseteq$ $\mathcal{X}$(AARC). Notice that for each base station b, the uncertainty set $\mathcal{Z}_b$ is a polytope given by a list of linear inequalities and an equality. Although the uncertainty set is in $\mathbb{R}^{K_b}$, because of the equality constraint, it is in a $K_b - 1$ dimensional subspace. Let n be the number of extreme points of $\mathcal{Z}_b$ and $\vec{\lambda}_1, \cdots, \vec{\lambda}_n$ be the extreme points. Then $\mathcal{Z}_b = \text{Conv}\{\vec{\lambda}_1, \cdots, \vec{\lambda}_n\}$. In the case of fixed recourse and a convex hull uncertainty set of a finite set, the ARC is given by the following LP.

$$\begin{array}{lll} \min & \vec{\alpha}, \vec{p}_{\lambda_1}, \cdots, \vec{p}_{\lambda_n} \sum_{l=1}^{L} \alpha_l & \\ \text{s.t.} & \rho_{bk} \leq \sum_{l=1}^{L} (p_{\lambda_j})^l_{bk} E_I[R^l_I | b, k] \quad \forall b, k \quad \forall j \\ & \sum_{k=1}^{K_b} (p_{\lambda_j})^l_{bk} \leq \alpha_l \quad \forall b, l \quad \forall j \\ & \sum_{l=1}^{L} \alpha_l \leq 1 \\ & (p_{\lambda_j})^l_{bk} \geq 0 \quad \forall b, k, l \quad \forall j \\ & \alpha_l \geq 0 \quad \forall l \end{array} \tag{16}$$

For each extreme point $\vec{\lambda}_j$, the second stage solution is given by $\vec{p}_{\lambda_j}$. For a point λ inside the uncertainty set, we know that it can be represented by a convex combination of K_b extreme points, $\vec{\lambda} = \sum_{j=1}^{K_b} c_j \vec{\lambda}_j$, since the uncertainty set is in a $K_b - 1$ dimensional subspace (Without loss of generality, let those points be the first K_b extreme points.) Moreover the solution for this point λ is given by a convex combination of the solutions of those extreme points, $\vec{p}_\lambda = \sum_{j=1}^{K_b} c_j \vec{p}_{\lambda_j}$.

Let's focus on a specific time fraction for base station b, customer class k under power profile l, p^l_{bk}. For this proof we further restrict our AARC policy so that it does not have a constant term. We have $p^l_{bk} = \sum_{m=1}^{K_b} \pi^l_{bkm}\lambda_{bm}$. Then by solving the following equation, we get an affine policy which agrees on every extreme point.

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1K_b} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2K_b} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{nK_b} \end{pmatrix} \cdot \begin{pmatrix} \pi^l_{bk1} \\ \pi^l_{bk2} \\ \vdots \\ \pi^l_{bkK_b} \end{pmatrix} = \begin{pmatrix} (p_{\lambda_1})^l_{bk} \\ (p_{\lambda_2})^l_{bk} \\ \vdots \\ (p_{\lambda_n})^l_{bk} \end{pmatrix},$$

where λ_{ij} is j^{th} coordinate of i^{th} extreme point. Since the uncertainty set is in $K_b - 1$ dimensional subspace, the following matrix has rank at most K_b, hence we can get an exact solution for $\pi^l_{bk1}, \pi^l_{bk2}, \cdots, \pi^l_{bkK_b}$ so that every pair of extreme point and its solution can be represented by our affine policy.

$$\left(\begin{array}{cccc|c} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1K_b} & (p_{\lambda_1})^l_{bk} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2K_b} & (p_{\lambda_2})^l_{bk} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \cdots & \lambda_{nK_b} & (p_{\lambda_n})^l_{bk} \end{array}\right)$$

Moreover for a general point λ inside the uncertainty set, the solution is given by

$$\begin{array}{rl} \vec{p}_\lambda & = \vec{\lambda} \cdot \vec{\pi} \\ & = (\sum_{j=1}^{K_b} c_j \lambda_j) \cdot \vec{\pi} \\ & = \sum_{j=1}^{K_b} c_j (\lambda_j \cdot \vec{\pi}) \\ & = \sum_{j=1}^{K_b} c_j \vec{p}_{\lambda_j} \end{array}$$

Therefore, any feasible solution of ARC can be represented by a feasible solution of AARC, which means $\mathcal{X}$(ARC)$\subseteq$ $\mathcal{X}$(AARC). ∎

3) Simulations and Results: For purposes of comparison, we evaluate the performance of our affine policy using the same simulation model of Rengarajan and de Veciana [1]. We consider three base stations facing each other in a hexagonal layout with radius 250m. A carrier frequency of 1GHz and a bandwidth of 10MHz are assumed. The base stations are assumed to be able to transmit at three different power levels: 0, 5 and 10W. The mean file size is 2MB.

We use different total arrival rates ranging from 0.5 to 2.2 and different protection levels, θ, restricting the uncertainty to an interval ranging from 0% to 40% of the nominal value. We use 100,000 customer samples to estimate the harmonic formula capacities and the mean delay. For each pair of total arrival rate and protection level, we randomly pick λ_{bk}'s in their bounds and compute the estimated delay 1,000 times to get average performance under the proposed uncertainties. Out of 1,000 experiments, we count the number of cases that the system becomes unstable and compute the average mean delay under the proposed uncertainties. As shown in Fig. 1(a),

at higher loads and higher uncertainty levels the number of unstable experiments is larger with the solution of the nominal optimization problem, i.e., $\theta = 0$. On the other hand, as shown in Fig. 1(b), with the solution of the AARC, the number of unstable experiments remains at zero until very high load, 2.0/sec. Note that even without fluctuations of arrival rates, the system is unstable around that point.

Fig. 2(a) and Fig. 2(b) show the average mean delay with the solution of the nominal optimization problem and the solution of the AARC respectively, at each load and uncertainty level. Since we cannot compute the mean delay of unstable systems, we draw average delay plots assigning the delay of an unstable experiment to be as large as the maximum delay over all delays of stable experiments. Thus the results we report are conservative, in the sense that since algorithms result in fewer unstable experiments, we are underreporting the decrease in average delay. At lower loads, even with high uncertainty levels the nominal solutions slightly outperform the AARC solutions. However, at higher loads, while the AARC solutions give acceptable low average mean delays, the nominal solutions give extremely high average mean delays even if the system is stable.

As we typically do not know precisely the uncertainty level in reality, we must balance the tradeoff between building in protection to uncertainty, and the loss of performance in the nominal setting, i.e., the cost of over-protection. To compare these factors, we pick three protection levels, 0% (nominal), 20% and 40%, and we consider the performance of these three solutions in different uncertainty regimes. Figure 3(a) shows the load at which stability breaks down for each solution, under large (40%) uncertainty. The nominal solution becomes unstable at a much lower average arrival rate than the 20%-protection and 40%-protection solutions. Interestingly, the 20%-protection solution remains stable for very heavy loads – essentially its stability performance is comparable to the 40%-protection solution. Figures 3(b,c,d) show the average mean delay (from simulation) of our three solutions. Figure 3(b) shows the delay curves when there is *no uncertainty* in arrival rates, i.e., the simulations are generated according to the nominal (and known) offered load. This shows the price of robustness. Indeed, as expected, the nominal solution outperforms both robust solutions giving lower delay – but the difference becomes pronounced only at very high loads. Meanwhile, Figures 3(c) and (d) illustrate the relatively quick deterioration of the nominal solution's delay performance under 20% and 40% uncertainty in the offered load. These results illustrate that the 20%-protection solution appears to have a low price of robustness, i.e., performance comparable to the nominal solution in the no-uncertainty regime, and yet captures most of the robustness properties of even the 40%-protection solution, outperforming the latter, except when both the load and the uncertainty level are high.

C. Uncertain Total Arrival Rate Model with Fixed Arrival Rates Ratio

1) Assumptions and an Adjustable Counterpart: In this section we consider a stochastic uncertainty model. We assume the arrival process is Poisson with rate λ and users arrive uniformly over the entire area. Thus for each class k and base station b, the *fraction* of arrival rate seen is fixed regardless of the change of the total arrival rate. We assume the total arrival rate, λ, changes according to a Markov process with drift towards the nominal rate. We discretize this process, approximating it via a Discrete Markov chain. We show that using this approximation preserves the convexity of the nominal problem.

We let Δ represent the quantization level for the arrival rate process, with probabilities $\bar{p}$ and $\bar{q}$ representing the drift away from and towards the nominal rate, respectively. The following is the transition matrix of the Markov chain we use:

$$\begin{array}{l}\forall i \geq 0,\\ Pr(\lambda(t+1) = \lambda^*(1+(i+1)\Delta) | \lambda(t) = \lambda^*(1+i\Delta)) = \bar{p}\\ Pr(\lambda(t+1) = \lambda^*(1+i\Delta) | \lambda(t) = \lambda^*(1+(i+1)\Delta)) = \bar{q}\\ \\ \forall i \leq 0,\\ Pr(\lambda(t+1) = \lambda^*(1+i\Delta) | \lambda(t) = \lambda^*(1+(i-1)\Delta)) = \bar{q}\\ Pr(\lambda(t+1) = \lambda^*(1+(i-1)\Delta) | \lambda(t) = \lambda^*(1+i\Delta)) = \bar{p},\end{array} \quad (17)$$

Although we know the full distribution of the arrival rate, we accept some error probability ϵ and truncate the distribution of the arrival rate into a finitely supported distribution. We do this because stability needs to be enforced deterministically, and hence must be enforced over the full support of the distribution. Let n be a number such that $Pr(\lambda \in [\lambda^*(1-n\Delta), \lambda^*(1+n\Delta)]) \geq 1-\epsilon$.

2) Affinely Adjustable Stochastic Counterpart of Objective: Computational experiments in [1] reveal that the solution obtained by minimizing delay indeed has improved delay performance over the solution obtained by maximizing capacity. In the stochastic formulation, the objective is an expected value over a discretely supported distribution, and hence becomes a sum of weighted variations. If the original objective is convex, then so is its expectation. Exploiting this fact and taking advantage of the knowledge of the distribution, we use the delay-minimizing objective rather than the capacity-optimizing objective we use in the robust formulation.

Then the objective of the stochastic optimization problem is as follows.

$$\begin{array}{l} E_\lambda\left[\sum_{b=1}^{N}\sum_{k=1}^{K_b}\frac{\frac{\rho_{bk}}{R_{bk}(\bar{p}_{bk}(\lambda))}}{1-\frac{\rho_{bk}}{R_{bk}(\bar{p}_{bk}(\lambda))}}\right] \\ = \sum_{i\in\{-n,n\}} Pr(\lambda = \lambda^*(1+i\Delta)) \\ \qquad \times \left(\frac{1}{\lambda}\sum_{b=1}^{N}\sum_{k=1}^{K_b}\frac{\frac{\rho_{bk}}{R_{bk}(\bar{p}_{bk}(\lambda))}}{1-\frac{\rho_{bk}}{R_{bk}(\bar{p}_{bk}(\lambda))}}\right).\end{array}$$

3) Affinely Adjustable Robust Counterpart of Constraints: Although we use a sampling method for the objective to preserve the convexity of the problem, we choose the similar robust optimization techniques we used in Section IV-B for the constraints. The stability constraints must be enforced *deterministically*. That is, we want to make the solution feasible for every realization of the arrival rate, hence the constraints must remain feasible for all arrival rates in the support of the truncated distribution, not only for the sample points on

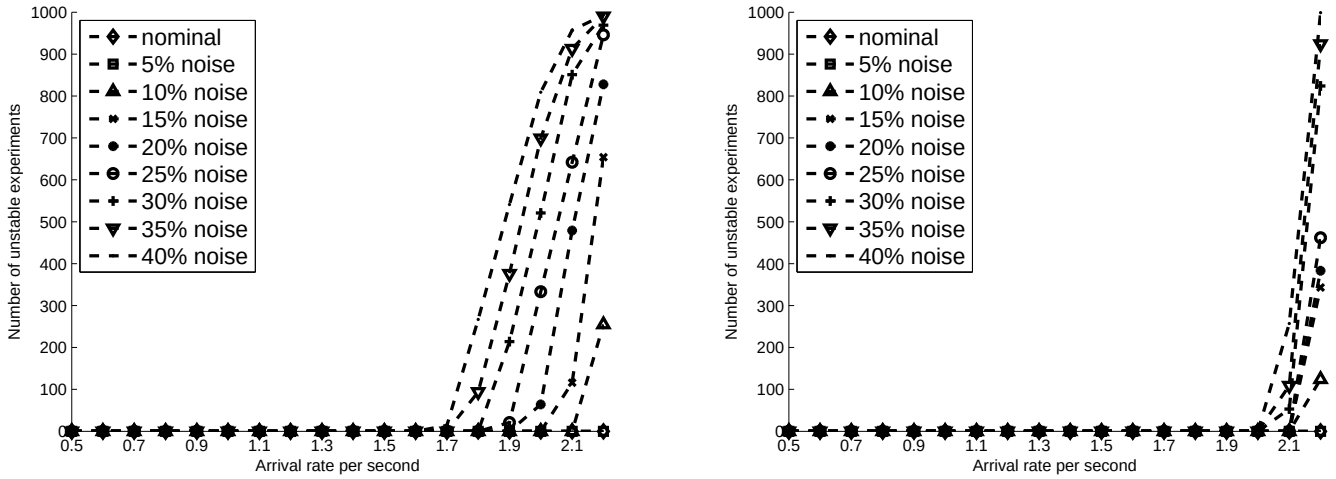

Fig. 1. AARC with capacity optimizing and affine policy on *fixed total arrival rate* model : Number of unstable experiments out of 1000 simulations against different actual uncertainty levels ranging from 0% to 40%: (a) the nominal solution, (b) each AARC solution runs against its predicted uncertainty level.

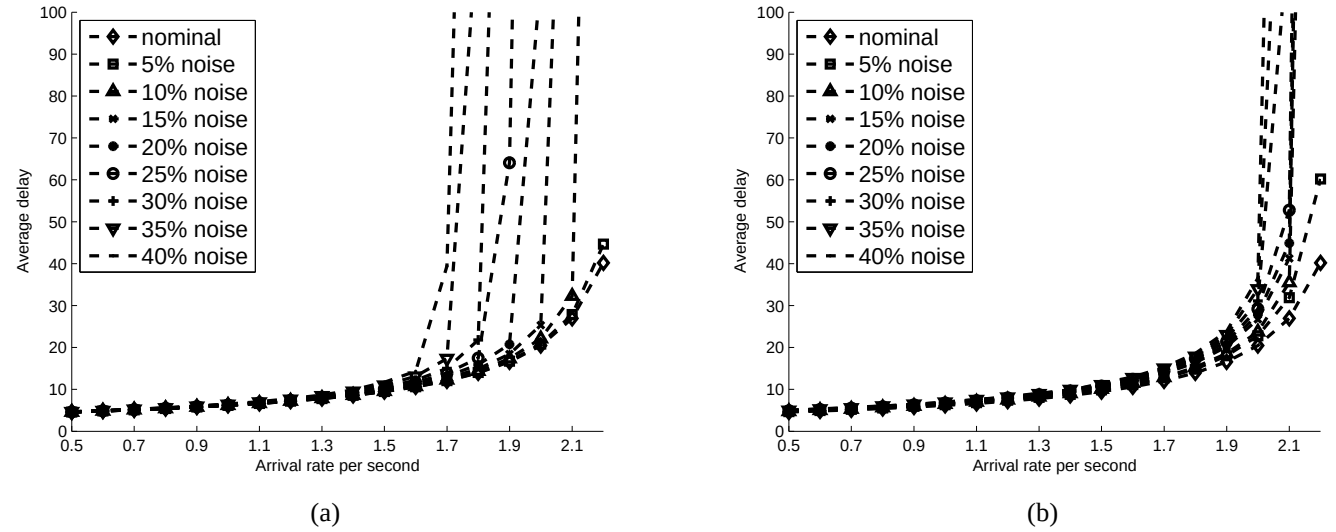

Fig. 2. AARC with capacity optimizing and affine policy on *fixed total arrival rate* model : Average file transfer delay against different actual uncertainty levels ranging from 0% to 40%: (a) the nominal solution, (b) each AARC solution runs against its predicted uncertainty level.

the discretized distribution. In the truncated distribution, the support of the total arrival rate, λ, is $[\lambda^*(1-n\Delta), \lambda^*(1+n\Delta)]$. Next, we restrict $p_{bk}^{\ l}$ to be an affine function of λ, i.e., $p_{bk}^{\ l} = \pi_{bk0}^{\ l} + \pi_{bk1}^{\ l}\lambda$. Let $\theta = n\Delta$ and $Y = [\lambda^*(1-\theta), \lambda^*(1+\theta)]$. Then as we discussed in Section IV-B1, θ represents the conservatism of the solution and Y is the support of the uncertainty set. We can control θ by adjusting the truncation error ϵ. If we allow smaller truncation error, then the support of the uncertainty set gets larger, i.e., n gets larger, hence the conservatism level θ becomes higher. Let $\tilde{\lambda}_{bk}$ be the fraction of arrival rate of user class k in base station b. Then $\lambda_{bk} = \tilde{\lambda}_{bk}\lambda$, and the resulting optimization problem is as follows.

$$
\begin{array}{ll}
\min & \vec{\alpha},\vec{\pi} \sum_{i\in\{-n,n\}} Pr(\lambda = \lambda^*(1+i\Delta)) \\
& \quad \times \left(\frac{1}{\lambda} \sum_{b=1}^{N} \sum_{k=1}^{K_b} \frac{\frac{\rho_{bk}}{R_{bk}(\vec{p}_{bk}(\lambda))}}{1-\frac{\rho_{bk}}{R_{bk}(\vec{p}_{bk}(\lambda))}} \right) \\
\text{s.t.} & \tilde{\lambda}_{bk}\lambda\overline{F}_{bk} \le \sum_{l=1}^{L}(\pi_{bk0}^{\ l} + \pi_{bk1}^{\ l}\lambda) E_I[R_I^l|b,k] \\
& \qquad\qquad\qquad\qquad \forall b,k \quad \forall\lambda \in Y \\
& \sum_{k=1}^{K_b}(\pi_{bk0}^{\ l} + \pi_{bk1}^{\ l}\lambda) \le \alpha_l \qquad \forall b,l \quad \forall\lambda \in Y \\
& \sum_{l=1}^{L} \alpha_l \le 1 \\
& \pi_{bk0}^{\ l} + \pi_{bk1}^{\ l}\lambda \ge 0 \qquad \forall b,k,l \quad \forall\lambda \in Y \\
& \alpha_l \ge 0 \qquad \forall l
\end{array}
\tag{18}
$$

Since we use an arithmetic mean approximation for R_{bk}, and since $\vec{p}_{bk}(\lambda)$ is linear in $\vec{\pi}$, R_{bk} is linear in $\vec{\pi}$. The term $\frac{\frac{\rho_{bk}}{R_{bk}(\vec{p}_{bk}(\lambda))}}{1-\frac{\rho_{bk}}{R_{bk}(\vec{p}_{bk}(\lambda))}}$ in the objective function is convex in $\vec{\pi}$

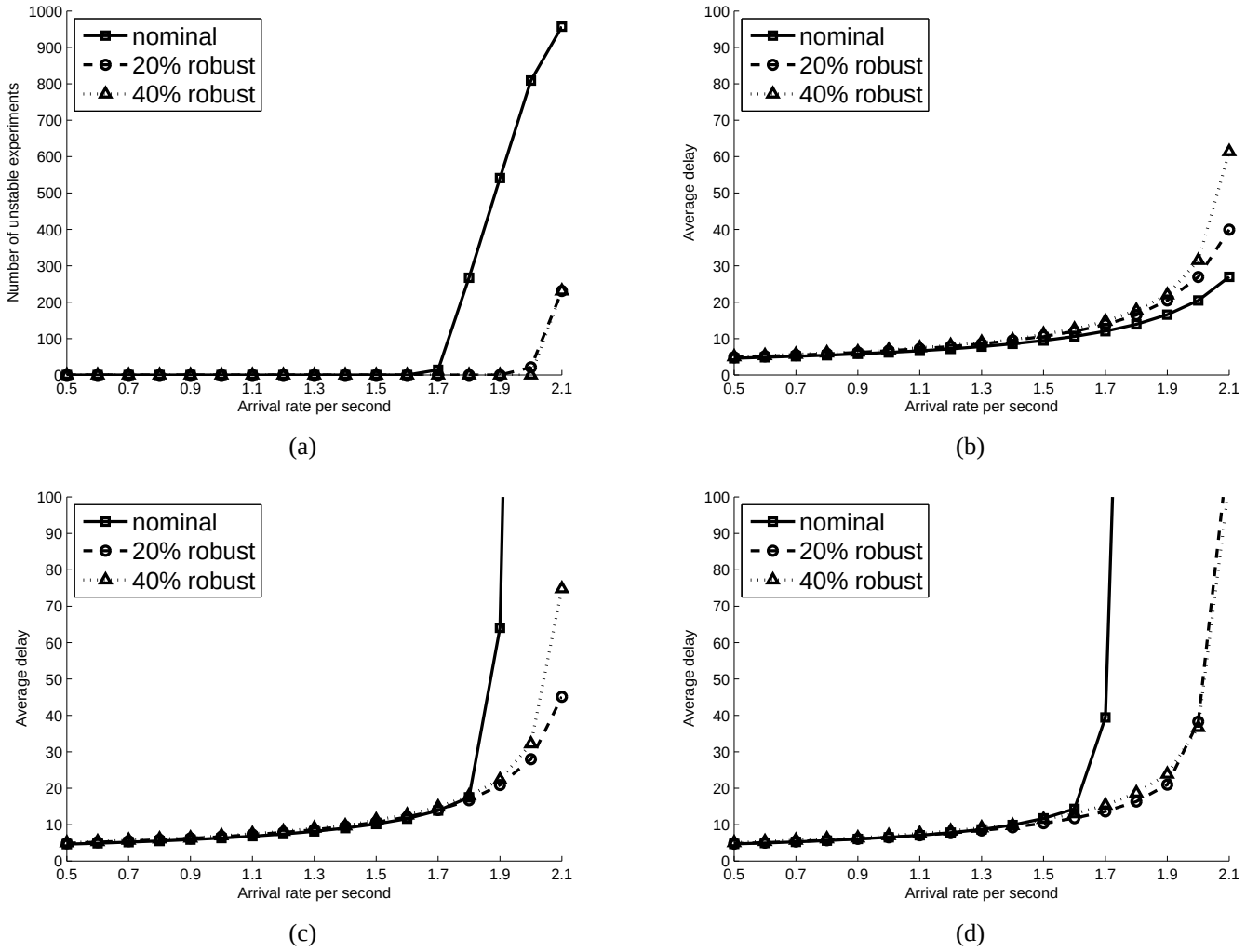

Fig. 3. AARC with capacity optimizing and affine policy on *fixed total arrival rate* model : (a) Number of unstable experiments out of 1000 simulations against 40% uncertainty level of offered loads, (b), (c) and (d) Average file transfer delay : (b) at the nominal offered loads, (c) against 20% uncertainty level of offered loads, (d) against 40% uncertainty level of offered loads.

if $\frac{\rho_{bk}}{R_{bk}(\vec{p}_{bk}(\lambda))} \leq 1$ ([1]) and indeed, this is the stability condition, and hence enforced for every λ in the support of its truncated distribution. Moreover, the infinite constraints ("$\forall \lambda \in Y$") can be transformed into a finite collection of linear constraints, again by employing a duality argument as in the previous section. Therefore this problem is a convex optimization problem with linear constraints.

4) Simulations and Results: We evaluate the performance of our optimization using the same simulation model of [1], and as before, using stability and delay as our evaluation metrics. We use different nominal arrival rates ranging from 0.5 to 2.2 and different truncation error, ϵ ranging from 1% to 20%. As we see in Section IV-C3, *lower truncation error means higher protection level*. We use 100,000 customer samples to estimate the harmonic formula capacities and the mean delay. At each simulation, we choose an arrival rate randomly from the stationary distribution of our Markov chain model. For each pair of nominal arrival rate and error probability, we run 1,000 experiments. We count the number of cases that the system is unstable and compute the average mean delay under the Markov chain model. We use the transition matrix (17) with $\bar{p} = 1/3$, $\bar{q} = 2/3$, and $\Delta = 6\%$.

As shown in Fig. 4(a), under nominal arrival rates, our uncertainty-protected solutions perform comparably to the nominal solution, hence the price of robustness is very low in this model.

Fig. 5(a) shows the number of unstable experiments for each solution and Fig. 5(b) shows the average delay under the uncertain arrival rates process changing acording to the Markov process described earlier. At higher loads, the number of unstable experiments is larger with the nominal solution. But, even the 20% truncation error model shows better results than the nominal solution in terms of stability and optimality. This is because the distribution of the uncertainty is concentrated around its mean. Even the truncated distribution with 20% error captures the original distribution well. The original distribution of the Markov chain with transition matrix (17) and the truncated distributions under different truncation errors are shown in Fig. 4(b). On the other hand, the constraints with less than 20% truncation error models (i.e., bigger θ) are overly

protective, and the conservativeness of the solution results in infeasibility at higher loads.

D. Revisiting the Uncertain Arrival Rates Ratio Model with Fixed Total Arrival Rate

Now we consider anew the *fixed total arrival rate* model of Section IV-B. Recall that there, although minimizing delay directly is more effective than maximizing capacity (as demonstrated by empirical results – see [1] for details), we have used the capacity maximization because the delay minimization is non-linear, and thus its AARC is not computationally tractable [9]. Tractability requires either convexity in the uncertain variables, or concavity along with an uncertainty set with a small set of extreme points. Neither of these hold for our model. To overcome this difficulty, in this section we use the stochastic optimization paradigm, following Section IV-C. First we briefly state our method and formulate the optimization problem. Then we show that the resulting problem is computationally tractable. Finally, we give some computational results.

1) Piecewise Linear Adaptability: In the AARC approach, we try to compute the optimal affine adaptable solution, however, as discussed above, this approach is intractable in this setting. Instead, we take advantage of the low dimensionality of the uncertainty set, and choose a finite collection of representative points. For each point, we choose the optimal value for the adaptable variables, and then extend to the full uncertainty set by interpolating between these points using bilinear functions.

The uncertainty set is determined by simple range constraints and one equality constraint, hence it is easy to find the extreme points of the uncertainty set. We use points uniformly distributed on the grid whose end points are those extreme points of the uncertainty set. This procedure is similar to the discretizing distribution method we use in Section IV-C. However, the uncertainty set is not one-dimensional as in the fixed ratio model, but rather has dimension equal to the number of customer classes. The number of points in the grid grows exponentially in the dimension. Therefore it is necessary to use a sparse grid, and use bilinearity to interpolate between the grid points, as explained below.

The grid effectively selects a finite number of representative uncertainty realizations from the uncertainty set. We enumerate the points on the grid as $\{\lambda_1, \ldots, \lambda_M\}$. We select a single power profile variable $\vec{\alpha}$ (first stage solution, that cannot depend on the realization of the uncertainty) and M customer profile variables, $\{\vec{p}_{\lambda_1}, \ldots, \vec{p}_{\lambda_M}\}$, corresponding to each of the M points on the grid, i.e., each of the realizations of the uncertainty. Assuming a uniform distribution over the M points $\{\lambda_1, \ldots, \lambda_M\}$, we choose $\vec{\alpha}, \{\vec{p}_{\lambda_1}, \ldots, \vec{p}_{\lambda_M}\}$ in order to minimize the mean delay over the points on the grid, and so that the power-profile/customer-profile pair $(\vec{\alpha}, \vec{p}_{\lambda_k})$ is feasible for the uncertainty realization λ_k, for $k = 1, \ldots, M$, where feaibility again means stability.

We accomplish this by solving the following optimization problem.

$$
\begin{array}{ll}
\min & \vec{\alpha}, \vec{p}_{\lambda_1}, \vec{p}_{\lambda_1}, \cdots, \vec{p}_{\lambda_M} \\
 & \sum_{j=1}^{M} \sum_{b=1}^{B} \sum_{k=1}^{K} -1 + 1 \Big/ \Big(1 - \frac{\rho_{bk}}{R_{bk}(\vec{p}_{\lambda_j})}\Big) \\
\text{s.t.} & (\lambda_j)_{bk} \overline{F}_{bk} \leq \sum_{l=1}^{L} (p_{\lambda_j})^l_{bk} E_I[R_I^l | b, k] \\
 & \qquad\qquad \forall b, k \ , \forall j \in \{1, \cdots, M\} \\
 & \sum_{k=1}^{K_b} (p_{\lambda_j})^l_{bk} \leq \alpha_l \qquad \forall b, l \ , \forall j \in \{1, \cdots, M\} \\
 & \sum_{l=1}^{L} \alpha_l \leq 1 \\
 & \vec{p}_{\lambda_j} \geq 0 \qquad\qquad \forall b, k, l \ , \forall j \in \{1, \cdots, M\} \\
 & \alpha_l \geq 0 \qquad\qquad \forall l
\end{array}
\tag{19}
$$

Now, if the realized uncertainty is some λ_k, we have already computed the optimal adaptable policy, namely, $\vec{p}_{\lambda_k}$. For a realized point which is not on the grid (as will typically be the case), we use bilinear interpolation using the solutions of sample points on the grid. If the realized point is outside the uncertainty set, we use the solution of nearest boundary sample point. This way, we obtain an approximation of the optimal second stage policy of the stochastic optimization problem, for the entire uncertainty. Since the optimal adaptable solution can be shown to be continuous, successively refining the grid allows arbitrary approximation of the optimal solution, although this comes at an increased computational cost, as M grows.

2) Simulations and Results: We use the same simulation model and uncertainty model of Section IV-B3. In order to approximate the uncertainty set, we use 5 by 5 size grids.

Fig. 6 shows comparisons of the performances of the AARC solution with capacity optimizing schedule we've obtained in Section IV-B and the stochastic solution with delay minimizing schedule of this section. One comparison without uncertainty in arrival rate(Fig. 6(a)) and the other with high uncertainty(Fig. 6(b)). As shown in the figures, the delay minimizing schedule outperforms the capacity optimizing schedule as we expected. The performance gap increases as total arrival rate increases.

Next we compare the performances of the nominal problem solution and the solutions of stochastic problems with 20% and 40% protection levels to see the cost of over-protection. As shown in Fig. 7(a), the cost is negligible. Then we compare the performances of those three solutions under large (40%) uncertainty to see the benefit of robustness. Fig. 7(b) shows a big performance gap between the original solution and the approximated stochastic solution with 40% protection level at higher loads.

V. DISCUSSION, CONCLUSION, AND FUTURE WORK

We proposed several different approaches that attempt to make the solution of the system level coordination optimization problem robust to the variations of offered loads under different models of uncertain data. In the case that each offered load fluctuates individually but the sum of variations is zero, we first used two-stage robust optimization with affine second-stage decisions, obtaining tractable optimization formulations to obtain solutions robust to variations in the offered load. Later we used approximated stochastic optimization with

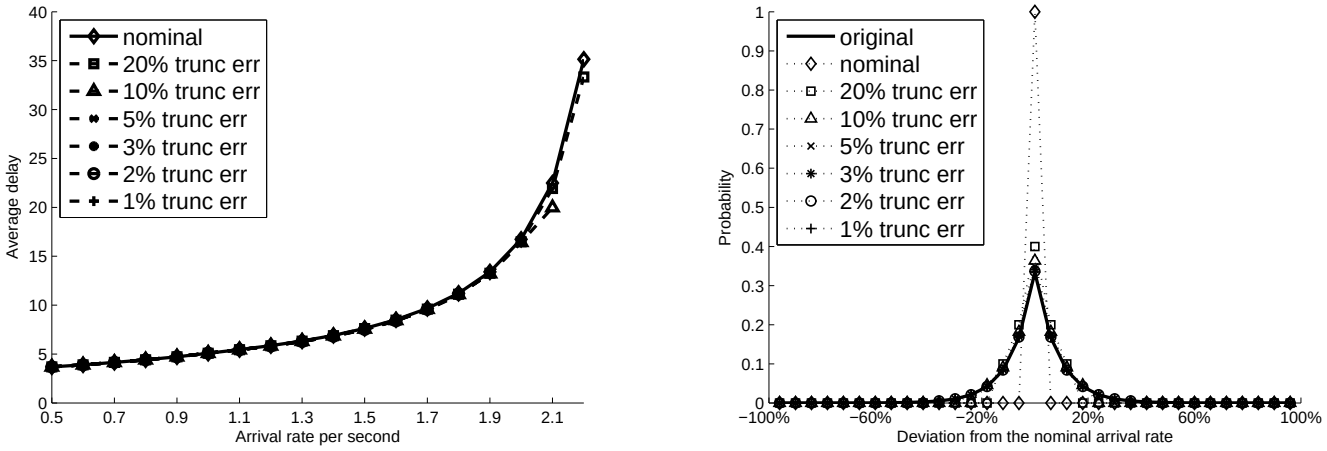

Fig. 4. (a) Stochastic problem with delay minimizing and affine policy on *fixed arrival rates ratio* model : Number of unstable experiments out of 1000 simulations at the nominal arrival rates, (b) Distribution of the uncertain arrival rate: Original distribution of the Markov chain v.s. Truncated distributions under different truncation errors.

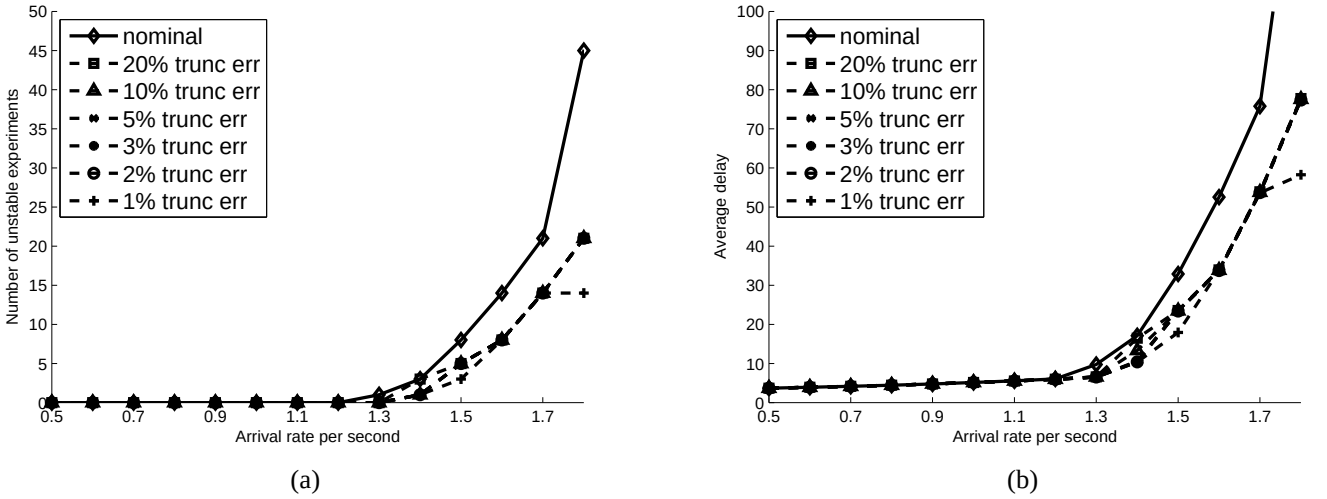

Fig. 5. Stochastic problem with delay minimizing and affine policy on *fixed arrival rates ratio* model : against the Discrete Markov chain uncertainty model of arrival rates with transition matrix (17) : (a) Number of unstable experiments, (b) Average file transfer delay.

sample points and interpolation. We also considered variation in the total arrival rate. There, we combined the stochastic optimization and the robust optimization paradigms, again obtaining solutions that remain stable under heavy loads, and get better average performance. In our simulation results, we have shown that nominal solutions are vulnerable to the fluctuations of the offered loads while properly tuned robust solutions capture the best of both worlds: resilience to uncertainty, with good performance even under the nominal setting.

Resilience to load variation could potentially help reduce coordination and hence communication requirements, without severely compromising the performance. Understanding the tradeoffs involved, between the benefits and costs of more frequent coordination is a key step towards understanding the viability of implementation of such a system-level optimization approach to interference mitigation, and is a topic of future work.

An issue we have not addressed here, and the subject of future work, is to treat the stochastic variation in the customer arrival process. Particularly in the low-load regime, this could result in empty customer classes, allowing base stations to (briefly) turn off, thus increasing the rates observed by customers of neighboring base stations. Optimizing coordination schemes to take advantage of this effect is a natural domain for multi-stage optimization models, although some considerable challenges stand in the way of immediate extensions of the methods presented here.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Balaji Rengarajan and Gustavo de Veciana for stimulating discussion on the topic, as well as for sharing past results and in particular simulation results, that made comparison on an even footing possible.

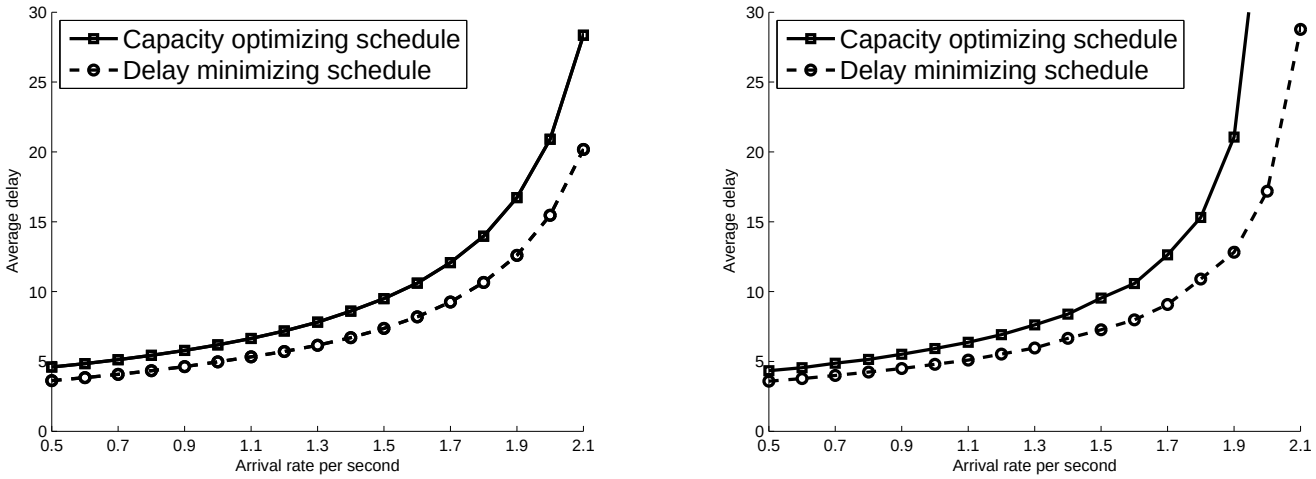

Fig. 6. *Fixed total arrival rate* model : (a) Average file transfer delay at the nominal offered loads with nominal solutions of capacity optimizing schedule and delay minimizing schedule, (b) Average file transfer delay against 40% uncertainty level of offered loads with 40% protection level solutions of capacity optimizing schedule and delay minimizing schedule.

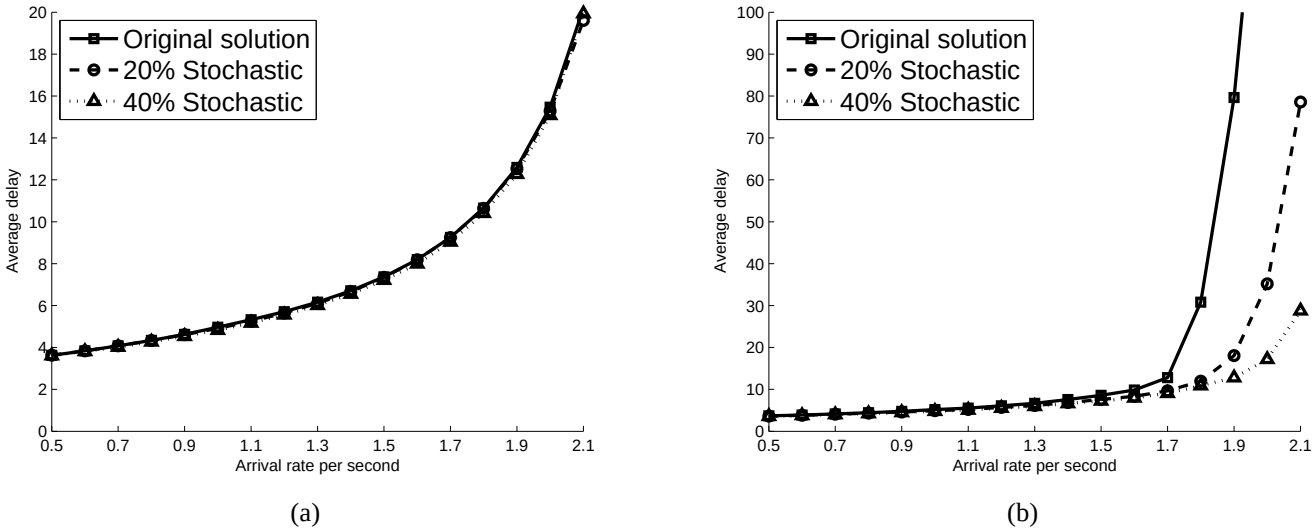

Fig. 7. Stochastic problem with delay minimizing and bilinear policy on *fixed total arrival rate* model : (a) Average file transfer delay at the nominal offered loads with different solutions (nominal, 20% protection level solution and 40& protection level solution), (b) Average file transfer delay against 40% uncertainty level of offered loads with different solutions.

REFERENCES

[1] B. Rengarajan and G. de Veciana, "Architecture and abstractions for environment and traffic aware system-level coordination of wireless networks: The downlink case," in *Proc. IEEE INFOCOM*, Apr 2008, pp. 1–9.

[2] S. Grandhi, R. Yates, and D. Goodman, "Resource allocation for cellular radio systems," *Vehicular Technology, IEEE Transactions on*, vol. 46, no. 3, pp. 581–587, Aug 1997.

[3] T. Fong, P. Henry, K. Leung, X. Qiu, and N. Shankaranarayanan, "Radio resource allocation in fixed broadband wireless networks," *Communications, IEEE Transactions on*, vol. 46, no. 6, pp. 806–818, Jun 1998.

[4] W. Rhee and J. Cioffi, "Increase in capacity of multiuser ofdm system using dynamic subchannel allocation," *Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo. 2000 IEEE 51st*, vol. 2, pp. 1085–1089 vol.2, 2000.

[5] D. Julian, M. Chiang, and D. O'Neill, "Robust and qos constrained optimization of power control in wireless cellular networks," *Vehicular Technology Conference, 2001. VTC 2001 Fall. IEEE VTS 54th*, vol. 3, pp. 1932–1936 vol.3, 2001.

[6] A. Atamturk and M. Zhang, "Two-Stage Robust Network Flow and Design Under Demand Uncertainty," *Operational Research*, vol. 55, no. 4, pp. 662–673, 2007.

[7] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski, "Adjustable robust solutions of uncertain linear programs," *Mathematical Programming*, vol. 99, pp. 351–376, 2004.

[8] A. Ben-Tal, S. Boyd, and A. Nemirovski, "Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems," *Mathematical Programming*, vol. 107, pp. 63–89, 2006.

[9] A. Ben-Tal and A. Nemirovski, "Robust solutions of linear programming problems contaminated with uncertain data," *Mathematical Programming*, vol. 88, pp. 411–424, 2000.

[10] D. Bertsimas and M. Sim, "Price of robustness," *Operational Research*, vol. 52, no. 1, pp. 35–53, Jan-Feb 2004.

[11] J. Birge and F. Louveaux, *Introduction to stochastic program*. New York: Springer, 1997.