Efficient Nonlinear Optimizations of Queuing Systems

Mung Chiang, Arak Sutivong, and Stephen Boyd
Electrical Engineering Department, Stanford University, CA 94305

Abstract— We present a systematic treatment of efficient non-linear optimizations of queuing systems. The suite of formulations uses the computational tool of convex optimization, with fast polynomial time algorithms to obtain the global optimum for these nonlinear problems under various constraints. We first show convexity structures of several queuing systems, including some surprising transition patterns, followed by formulating and showing numerical examples of several convex performance optimizations for both single queues and queuing networks. Blocking probability minimization and service rate allocation through the effective bandwidth approach is also presented.

I. INTRODUCTION

Queuing systems form a fundamental part for different types of networks, including computer multiprocessor networks and communications data networks. Queuing systems are also an integral part of various network elements, such as the input and output buffers of a packet switch. We often would like to optimize some performance metrics of queuing systems, for example, buffer occupancy, overall delay, jittering, workload, and probabilities of certain states. In a network of queues, we may also have multiple conflicting objectives that need to be optimally balanced. However, optimizing the performance of even simple queues like the $M/M/m/m$ queue is in general a difficult problem because of the nonlinearity of the performance metrics as functions of the arrival and service rates. Nonlinear optimization in general takes running time that scales exponentially with the problem size.

We show how convexity properties of queuing systems can be used to turn some of these intractable problems into polynomial time solvable ones. By using the tool of convex optimization, and in particular, geometric programming, we provide a suite of formulations to efficiently optimize the performance of queuing systems under Quality of Service (QoS) and fairness constraints, first for single queues in section III, then for blocking probability minimization and service rate allocation through the effective bandwidth approach in section IV, then for networks of queues in section V, and for optimal feedback control in simple queuing networks in section V. The distinguishing characteristics of the formulations in this paper is that they are nonlinear problems that can be solved as easily as linear problems by using convex optimization.

II. CONVEX OPTIMIZATION AND GEOMETRIC PROGRAMMING

Convex optimization refers to minimizing a convex objective function subject to upper bound inequalities on convex constraint functions. The objective function can be generalized to be vector-valued, where the minimization is with respect to a convex cone. These convex multiple-objective optimizations are useful for tradeoff analysis, and the notion of optimality now becomes Pareto optimality [1].

Convex optimization problems can be easy to solve, both in theory and in practice. Theoretically, showing an optimization problem to be a strictly convex problem proves that there is a unique global optimal solution, and leads to performance bounds and sensitivity analysis through the dual problem. Practically, when put in the right form, convex optimization can be globally solved by fast polynomial time algorithms [9]. It also gives a good starting point to develop even simpler heuristics and establishes the optimal benchmark to compare heuristics with.

There is a particular type of convex optimization used in sections III, IV and V called geometric programming [1], [4], which has also been applied to solve other network resource allocation problems [6]. First, we have

Definition 1: A monomial is a function $f : \mathbf{R}^n \to \mathbf{R}$, where the domain contains all real vectors with positive components, and constants $c \geq 0, a_i \in \mathbf{R}$:

$$f(x) = c x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}. \tag{1}$$

Definition 2: A posynomial is a sum of monomials $f(x) = \sum_k c_k x_1^{a_{1k}} x_2^{a_{2k}} \cdots x_n^{a_{nk}}$.

Geometric programming is an optimization problem in the following form:

$$
\begin{aligned}
\text{minimize} \quad & f_0(x) \\
\text{subject to} \quad & f_i(x) \leq 1, \\
& h_j(x) = 1
\end{aligned}
\tag{2}
$$

where f_0 and f_i are posynomials and h_j are monomials. Geometric programming in the above form is not a convex optimization problem. However, with a change of variables: $y_i = \log x_i$ and $b_{ik} = \log c_{ik}$, it can be shown that the reformulated problem is a convex optimization problem [1].

III. CONVEX OPTIMIZATIONS OF SINGLE QUEUES

A. Optimizing for average delay and queue occupancy

We start the suite of convex optimization formulations with a simple example of minimizing the service load of a $M/M/1$ queue with constraints on average queuing delay W, total delay D, and queue occupancy Q:

Proposition 1: The following nonlinear optimization is a geometric program, and therefore can be turned into a convex optimization and efficiently solved for its global optimum:

$$
\begin{array}{ll}
\text{minimize} & \frac{\mu}{\lambda} \\
\text{subject to} & W \leq W_{max}, \\
& D \leq D_{max}, \\
& Q \leq Q_{max}, \\
& \lambda \geq \lambda_{min}, \\
& \mu \leq \mu_{max}
\end{array}
\tag{3}
$$

where the optimization variables are the arrival rate λ and the service rate μ. The constant parameters are the performance upper bounds W_{max}, D_{max} and Q_{max}, and practical constraints on the maximum service rate μ_{max} of the queue that cannot be exceeded, and the minimum incoming traffic rate λ_{min} that must be supported. The objective is to minimize the service load. We can also show that even a joint optimization over both (λ, μ) and $(W_{max}, D_{max}, Q_{max})$ is still a geometric program.

The above formulation can be extended to a Markovian queuing system with N queues sharing a pool of service rate bounded by μ_{max} (for example, connected to a common outgoing link). The arrival rate to be supported for each individual queue i is bounded by $\lambda_{i,min}$. There are delay and queue occupancy bounds $W_{i,max}$, $D_{i,max}$ and $Q_{i,max}$ for each queue i. The objective now becomes minimizing a weighted sum of the service loads for all the queues:

Corollary 1: The following nonlinear optimization is a geometric program:

$$
\begin{array}{ll}
\text{minimize} & \sum_{i=1}^{N} \alpha_i \frac{\mu_i}{\lambda_i} \\
\text{subject to} & W_i \leq W_{i,max}, \\
& D_i \leq D_{i,max}, \\
& Q_i \leq Q_{i,max}, \\
& \lambda_i \geq \lambda_{i,min}, \\
& \sum_{i=1}^{N} \mu_i \leq \mu_{max}
\end{array}
\tag{4}
$$

where the optimization variables are the arrival rates λ_i and the service rates μ_i.

A simple numerical example for $N = 2$ with weights $\alpha_1 = 1, \alpha_2 = 2$ is summarized as follows. If we set the delay and queue occupancy constraints as $Q_{1,max} = 4, Q_{2,max} = 5, W_{1,max} = 2.5, W_{2,max} = 3, D_{1,max} = 2, D_{2,max} = 2$, and service and arrival rate constraints as $\lambda_{1,min} = 0.5, \lambda_{2,min} = 0.8, \mu_{max} = 3$, geometric programming gives the optimizers: $\mu_1^* = 1.328, \mu_2^* = 1.672, \lambda_1^* = 0.828, \lambda_2^* = 1.172$ and the optimized objective value is 4.457.

B. Optimizing $M/M/m/m$ queues

We now optimize specific queue occupancy probabilities by first considering an $M/M/m/m$ queue. The steady state probability of state k is given by $p_k = \frac{(\frac{\lambda}{\mu})^k \frac{1}{k!}}{\sum_{i=0}^{m}(\frac{\lambda}{\mu})^i \frac{1}{i!}}$. In many applications of queuing systems to network design, we would like to maximize the probability of a particular desirable state, without making the probabilities of other states too small. For example, we may want to design a telephone call service center so as to maximize the probability that a particular number of telephone lines (e.g., 90%) are in use at any given time. We also want to jointly optimize the fairness parameters C_j that bounds p_j.

Proposition 2: The following nonlinear optimization of $M/M/m/m$ queues is a geometric program:

$$
\begin{array}{ll}
\text{maximize} & p_k \\
\text{subject to} & p_j \geq C_j, \quad \forall j, \\
& C_j \geq C_{j,min}, \quad \forall j, \\
& \lambda \geq \lambda_{min}, \\
& \mu \leq \mu_{max}
\end{array}
\tag{5}
$$

where the optimization variables are λ, μ and C_j, and the constant parameters are λ_{min}, μ_{max} and the fairness constraints $C_{j,min}, j = 1, 2, \cdots, m$.

This geometric programming formulation can be extended to maximize the probability for the state with the lowest probability to enforce maxmin fairness. Similar formulations can be done for parallel $M/M/m/m$ queues and a general M/G queue.

C. Optimizing $M/M/1$ queues

We now turn to $M/M/1$ queues, where the convexity property is, surprisingly, more complicated than that of queues with finite buffer size. We first prove the convexity properties of the relevant quantities and then show the appropriate nonlinear optimization formulations. For $M/M/1$ queues, the state probability p_k can be viewed as a function of either the traffic load $\rho = \frac{\lambda}{\mu}$ or of λ and μ.

The state probability p_1 is always a concave function of ρ. However, there is an interesting transition from convexity to concavity as load increases for $p_k, k \geq 2$, derived from the second derivative test of convexity and shown in the following

Lemma 1: The state probability $p_k, k \geq 2$ is a convex function of load ρ if and only if $(k - 1) - (k + 1)\rho \geq 0$.

Therefore, there is a transition from convexity to concavity across the states in ascending order as load increases from $\frac{1}{3}$ to 1. In order for p_k to be convex in ρ for all k greater than or equal to a critical k_0, the traffic load ρ must be smaller than a critical $\rho_0(k_0)$. Numerically evaluating the above condition, we obtain the convexity transition curve shown in Figure 1, where the critical load ρ_0 can be read for any given k_0.

We now turn to the more useful design problem where we can vary the arrival rate and service rate independently, instead of
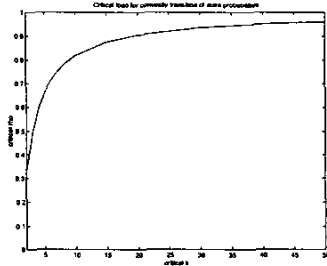
Fig. 1. Threshold loads for transition of $p_k(\rho)$ from convexity to concavity for a $M/M/1$ queue.

just their ratio ρ. Unlike $M/M/m/m$ queues where geometric programming can be used for optimizing over λ and μ, the state probabilities p_k of an $M/M/1$ queues are not in general convex functions of λ and μ. There is a similar, though more complicated, pattern of this transition when p_k is viewed as a function of two variables λ and μ. Interestingly enough, this transition pattern still only depends on ρ, as shown in the following

Lemma 2: The function p_k, $k \geq 2$, is convex in λ and μ if and only if $(k^2+k) - (k^2+k)\rho + (k^2-k)\rho^2 - (k^2+k+2)\rho^3 \geq 0$. This lemma leads to the following

Proposition 3: The following nonlinear optimization of $M/M/1$ queues is a convex optimization problem:

$$
\begin{aligned}
\text{minimize} \quad & p_k \\
\text{subject to} \quad & p_j \leq C_j, \quad j > k, \\
& \mu \leq \mu_{max}, \\
& \lambda \geq \lambda_{min}, \\
& \rho < \rho_0(k)
\end{aligned}
\tag{6}
$$

where $0 \leq \rho_0(k) < 1$ solves the equation $(k^2 + k) - (k^2 + k)\rho + (k^2 - k)\rho^2 - (k^2 + k + 2)\rho^3 = 0$. The optimization variables are λ and μ, and the constant parameters are $C_j, \lambda_{min}, \mu_{max}$ and k.

IV. OPTIMIZATIONS WITH BUFFER OVERFLOW CONSTRAINTS THROUGH EFFECTIVE BANDWIDTH

One approach to study the buffer overflow probability is through the blocking probability of an $M/M/1/B$ queue with a fixed buffer of size B:

$$
p_B = \frac{(\frac{\lambda}{\mu})^B \frac{1}{B!}}{\sum_{i=0}^{B} (\frac{\lambda}{\mu})^i \frac{1}{i!}}.
$$

Therefore, minimizing p_B is equivalent to maximizing a posynomial of λ and μ, which is in turn equivalent to maximizing a convex function. Therefore, it is not a convex optimization problem. One possible heuristics is to use geometric programming to maximize the probability of some state k, $k < B$, subject to lower bounds on p_j for all other $j < B$.

Since $p_B = 1 - \sum_{i=0}^{B-1} p_i$, this heuristics essentially minimizes the blocking probability. It is also known that due to the superadditive effect of buffer size on p_B, allocating a fixed buffer space among several queues to minimize the overall blocking probability is also a convex optimization.

An alternative way to characterize buffer overflow is through the large deviation approach, where the blocking probability is guaranteed statistically: for a connection X with a prescribed service rate R in the queue, we would like to ensure that the probability of overflow (receiving more than R bps from X) over a time scale of t is exponentially small:

$$
\text{Prob}\left\{ \sum_{i=1}^{t} X(i) \geq R \right\} \leq \exp(-sR) \tag{7}
$$

where $s \geq 0$ is the undersubscription factor. Smaller s implies more aggressive statistical multiplexing of multiple connections to one queue. This number R is called the effective bandwidth EB of X (as first proposed in [5], used in many papers since, and nicely reviewed in [7]).

Using the Chernoff bound, the effective bandwidth is given by

$$
\text{EB}(X) = \frac{1}{st} \log \mathbf{E}\left[\exp(sX)\right], \tag{8}
$$

and in practice, the expectation is replaced by empirical data collected over a time period of $\tilde{t}$ that is much larger than the time scale factor t:

$$
\text{EB}(X) = \frac{1}{st} \log \left(\frac{t}{\tilde{t}} \sum_{i=1}^{\frac{\tilde{t}}{t}} \exp(sX(i)) \right)
$$

where $X(i)$ is the number of bits produced by connection X during the ith time slot.

We want to either minimize the assigned service rate $\text{EB}(X)$ subject to constraints that lower bound the traffic intensity $X(i)$ to be supported (i.e., exponentially small probability of overflow or blocking), or maximize the traffic intensity subject to constraints upper bounding the service rate that can be assigned to X. Both problems are geometric programs, and we focus on the first formulation for the rest of this section (since it is also connected with information theoretic channel capacity [3]), where we put various constraints (indexed by j and induced by the stochasticity of other connections sharing the queue buffer) on the minimal level of traffic intensity to be supported by $\text{EB}(X)$.

Proposition 4: The following problem of constrained buffer allocation through the effective bandwidth approach is a geometric program:

$$
\begin{aligned}
\text{minimize} \quad & \text{EB}(X) \\
\text{subject to} \quad & \sum_i P_{ij} X(i) \geq X_{min,j}, \quad \forall j
\end{aligned}
\tag{9}
$$

where the optimization variables are $X(i)$, and the constant parameters are P_{ij} and $X_{min,j}$.

An illustrative numerical example is summarized as follows. With $s = 0.5, t = 5$ms, we impose a set of 10 different constraints to specify the type of an arrival curve a queue should be able to support without blocking. The geometric programming solution returns the minimized effective bandwidth as $EB^*(X) = 1.7627$Mbps, and the envelope of supportable arrival curves is shown in Figure 2. Connections with arrival curves below this envelope will not cause buffer overflow or queue blocking with a probabilistic guarantee as in (7).
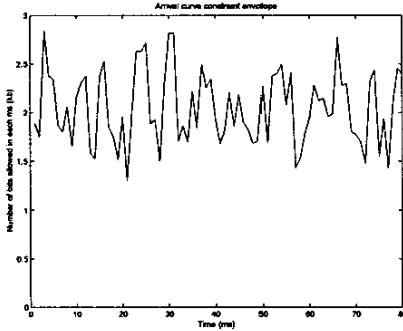

Fig. 2. The envelope of arrival curves supportable by $EB^*(X) = 1.7627$.

V. Convex Optimizations of Queuing Networks

In some queuing problems, a fixed number of customers or tasks circulate indefinitely in a closed network of queues. For example, some computer system models assume that at any given time a fixed number of programs occupy the resource. Such problems can be modelled by a closed queuing network consisting of K nodes, where each node k consists of m_k identical exponential servers, each with average service rate μ_k. There are always exactly N customers in the system. Once served at node k, a customer goes to node j with probability p_{kj}. Then for each node k, the average arrival rate to the node, λ_k, is given by $\lambda_k = \sum_{j=1}^{K} p_{kj}\lambda_j$.

The steady state probability that there are n_k customers in node k, for $k = 1, 2, \cdots, K$, is given by [8] (a closed network Jackson's theorem):

$$p(n_1, n_2, \cdots, n_K) = \frac{1}{G(K)} \prod_{k=1}^{K} \frac{\left(\frac{\lambda_k}{\mu_k}\right)^{n_k}}{\beta_k(n_k)},$$

where

$$\beta_k(n_k) = \begin{cases} n_k! & n_k \leq m_k \\ m_k! m_k^{n_k - m_k} & n_k > m_k \end{cases}$$

and the normalization constant $G(K)$ is given by

$$G(K) = \sum_{s} \prod_{k=1}^{N} \frac{\left(\frac{\lambda_k}{\mu_k}\right)^{n_k}}{\beta_k(n_k)}$$

where the summation is taken over all state vectors $(n_1, n_2, \cdots, n_K)$ satisfying $\sum_{k=1}^{N} n_k = N$. We have

Proposition 5: The nonlinear problem of maximizing the probability of state $(n_1 = n_1^*, \cdots, n_K = n_K^*)$ with $\sum_{k=1}^{N} n_k = N$, subject to fairness constraints on other states, is a geometric program:

maximize $p(n_1 = n_1^*, \cdots, n_K = n_K^*)$
subject to $p(n_1, \cdots, n_K) \geq$ Fairness constants,
$\mu_i \leq \mu_{i,max}$,   (10)
$\mu_i \geq \mu_{i,min}$,
$\sum_{k=1}^{K} m_k \mu_k \leq \mu_{total}$

where there is a constraint of the first type for each steady state probability $p(n_1, \cdots, n_K)$. The optimization variables are μ_k, and the constant parameters are $\mu_{i,min}$, $\mu_{i,max}$ and μ_{total}.

The above convex optimization problem can be viewed as a problem of resource (i.e., service capacity μ_k) allocation in a closed queuing network. The goal is to maximize the probability that the system is in a particular state subject to fairness constraints on other states and the limited system resource.

At first glance, it may seem that the above formulations can be readily extended to an open queuing network. However, because of a more complicated convexity structure of an open network (in particular, the steady state probability $p(n_1, n_2, \cdots, n_K)$ is neither concave nor convex in ρ_k), a similar formulation can be intractable for a general open network of queues. For a simple example, consider an open queuing network with two interconnected $M/M/1$ queues, each with an exponential service rate μ_k and an external arrival rate $\alpha_k, k = 1, 2$. After being served by queue k, a customer chooses to go to the other queue with probability p_k or leave the queuing system with probability $1 - p_k$. It can be shown that $p(n_1, n_2)$ is neither concave nor convex.

VI. Convex Optimizations of Feedback Control in Queuing Networks

In this section, we extend the convex optimization formulations to a particular type of queuing networks with feedback. Although the formulations are no longer in the special form of geometric program, we can still turn them into a general convex optimization problem.

As a first example, consider a simple network of queues shown in Figure 3.
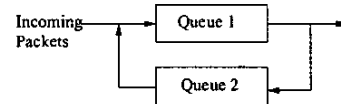

Fig. 3. A network of queues with feedback.

The incoming traffic to the overall system is i.i.d. $\sim$ Poisson(λ). With probability p_1, an incoming packet leaves the system after the feedforward queue 1, which has an exponential service time μ_1, and with probability $p_2 = 1 - p_1$, the packet is feed back through queue 2, which has an exponential service time μ_2. This queuing model can be used for

a variety of systems where we would like to process as many packets through the feedback loop as allowed under a delay constraint on the total time spent in the system. For example, in some optical packet switch architectures, the problem of wavelength contention can be solved by cycling packets not switched in a time slot through the buffer again. Clearly, there is a tradeoff between maximizing the feedback queue traffic load ρ_2 (or equivalently, minimizing the service load $\frac{1}{\rho_2}$) and minimizing the total time T spent in the system. We have

Proposition 6: The nonlinear optimization problem of minimizing both T and $\frac{1}{\rho_2}$ by varying p_2, subject to the following constraints: $\rho_1 < 1, \rho_2 < 1$, and $0 \le p_2 \le 1$ is a convex multi-objective optimization problem, where all the Pareto optimal solutions can be found through a scalarization technique as follows.

The problem of minimizing a weighted sum of T and $\frac{1}{\rho_2}$, subject to stability constraints for each individual queue, is a convex optimization problem in variable p_2:

$$
\begin{aligned}
\text{minimize} \quad & T + \alpha \frac{1}{\rho_2} \\
\text{subject to} \quad & \rho_1 < 1, \\
& \rho_2 < 1, \\
& p_2 \le 1, \\
& p_2 \ge 0
\end{aligned}
\tag{11}
$$

Therefore, the nonlinear problem of finding the best feedback parameter p_2^* and $p_1^* = 1 - p_2^*$ to minimize the total system time and maximize the feedback queue traffic load under individual queue stability constraints can be efficiently solved for the globally optimal solution.

We summarize a numerical example for the case of one feedback queue. Starting with the scalarized version, we first fix a weighting factor $\alpha = 0.5$ for $\lambda = 5, \mu_1 = 8$ and $\mu_2 = 8$. As shown in Figure 4, the convex optimization algorithm finds the global optimum value for the objective function as 3.933 through the optimal feedback probability $p_2^* = 0.2755$.
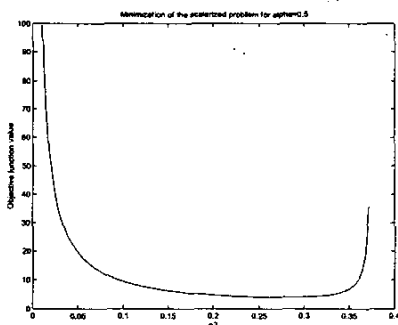

Fig. 4. Optimizing $T + \frac{\alpha}{\rho_2}$ over p_2 for a fixed α.

Now we solve the multi-objective problem as in Proposition 6 through scalarization of this convex Pareto optimization. Due to the convexity structure, by varying α we can obtain the entire tradeoff curve shown in Figure 5, where each point on the curve corresponds to the result of a convex scalar optimization

solution for a fixed α. Note that only points to the right of the Pareto optimality curve are achievable.
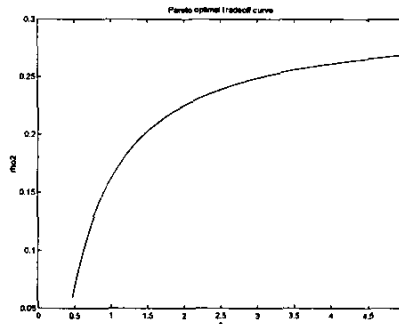

Fig. 5. Pareto optimality tradeoff curve as α varies.

With two parallel feedback queues, maximizing a weighted sum of feedback queue loads subject to upper bounds on the feedforward queue load is a convex optimization, so is minimizing the ratio of the feedforward load and the sum of feedback loads. However, due to more involved convexity structures of the queuing system, extending the above analysis to a general case with n feedback queues is not straightforward,

VII. CONCLUSIONS

Based on various results on convexity properties of queuing systems and computationally efficient algorithms for convex optimization, we present a suite of formulations to optimize the performance of single queues, networks of queues, and large deviation theoretic bounds on blocking probability minimization. These nonlinear performance optimizations can be carried out globally in polynomial time for network queuing systems under QoS and fairness constraints.

REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization* Stanford University EE 364 Course Reader 2001.

[2] C. S. Chang, "Stability, queue length and delay of deterministic and stochastic queuing networks," *IEEE Trans. Automatic Control,* vol. 39, pp. 913-931, 1994.

[3] M. Chiang and S. Boyd, "Shannon duality through Lagrange duality: efficient computation and free energy interpretations for channel capacity and rate distortion," *Proc. 40th Allerton Conference,* Oct. 2002.

[4] R. J. Duffin, E. L. Peterson, C. Zener, *Geometric Programming: Theory and Applications,* Wiley 1967.

[5] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Sel. Area Comm.,* pp.1598-1608, 1988.

[6] D. Julian, M. Chiang, D. O'Neill, and S. Boyd, "QoS and fairness constrained convex optimization of resource allocation in wireless cellular and ad hoc networks," *Proc. IEEE Infocom,* New York, June 2002.

[7] F. P. Kelly, "Notes on effective bandwidth," *Stochastic Networks: Theory and Applications,* pp.141-168, Oxford Unviersity Press, 1996.

[8] L. Kleinrock, *Queueing Systems, vol.I,* Wiley, 1974.

[9] Yu. Nesterov and A. Nemirovsky, *Interior Point Polynomial Method in Convex Programming,* SIAM 1994.