Title
A Convex Optimization Framework for Service Rate Allocation in Finite Communications Buffers

Journal
IEEE Communications Letters, 20(1)

ISSN
1089-7798

Authors
Ghamkhari, Mahdi
Mohsenian-Rad, Hamed

Publication Date
2016

DOI
10.1109/LCOMM.2015.2499740

Peer reviewed

A Convex Optimization Framework for Service Rate Allocation in Finite Communications Buffers

Mahdi Ghamkhari, *Student Member, IEEE* and Hamed Mohsenian-Rad, *Senior Member, IEEE*

***Abstract*—We study the convexity of loss probability in communications and networking optimization problems that involve finite buffers, where the arrival process has a *general* distribution. Examples of such problems include scheduling, energy management and revenue and cost optimization problems. To achieve a computationally tractable optimization framework, we propose to adjust an existing non-convex loss probability formula for G/D/1 queues to present a convex and even more accurate loss probability model. We then use empirical data and computer simulations to examine the performance of the proposed design.**

***Keywords*: Convex optimization, finite buffers, loss probability.**

I. Introduction

Loss Probability is commonly used as a metric to assess the performance of communication and networking systems that involve finite buffers [1], [2]. It is usually modeled as a function of service rate. A higher service rate results in a lower loss probability, which in turn improves quality-of-service (QoS). However, a higher service rate may increase the cost of service, e.g., due to using additional equipment or resources. Therefore, there is a *trade-off* between maximizing performance and minimizing cost in selecting service rate. This trade off can be systematically captured within an optimization framework where the service rate is the optimization variable. However, a major concern is whether the problem is convex and tractable. If an optimization problem that involves loss probability is not convex, as in [3]–[5], then inaccurate local optimization or time-consuming exhaustive search are needed to solve the problem. Prior studies have previously examined the convexity of loss probability models in certain queueing systems such as M/M/1/K queues, c.f. [6]. However, in this paper, our focus is on the steady state behavior of finite G/D/1 buffers, where the arrival process has a *general* distribution and the queue is first-in first-out (FIFO), non-preemptive and non-process sharing.

II. Example Service Rate Allocation Problems

A. Case 1: Maximum Profit Multi-Service Scheduling

Consider a communications, networking, or computation system with $N \geq 1$ finite buffers to admit N different service types. For each service type $i = 1, \ldots, N$, the arrival rate is modeled using its mean λ_i, variance σ_i^2, and auto-covariance $\rho_i(l)$, where l is the lag-time. Let $\mu_i \geq 0$ denote the rate at which the service requests of type i are handled. For a time interval of length T, let $q_i(\mu)$ denote the loss probability at queue i. Let I_i denote the number of service requests of type i that are handled within this time interval of interest. We have

$$E\{I_i\} = T\lambda_i \left(1 - q_i(\mu_i)\right). \tag{1}$$

Next, let $R_i(\cdot)$ denote the revenue function that indicates the revenue the server receives as a function of the total number of handled service requests of type i, c.f. [1], [2], [7]. The *per-time interval* revenue is calculated as $R_i(T\lambda_i(1 - q_i(\mu_i)))$. Similarly, let $C_i(\mu_i)$ denote the cost that incurs to the server due to handling the service requests of type i. The revenue and cost functions are assumed to be non-decreasing functions of their arguments. They are also concave and convex in their arguments, respectively. To maximize the total profit in the system, we need to solve the following optimization problem:

$$\underset{\mu_i \geq 0}{\textbf{Maximize}} \;\; \textstyle\sum_{i=1}^{N} R_i(T\lambda_i(1 - q_i(\mu_i))) - \sum_{i=1}^{N} C_i(\mu_i) \tag{2}$$

Using the composition rules [8, p. 85], we can verify that the above problem is convex as long as the loss probability model $q_i(\mu_i)$ is convex in μ_i for every service type i.

B. Case 2: Stochastic Service Rate Optimization

In practice, there can be *uncertainties* even with respect to the exact statistical characteristics of the arrival process in communications systems, e.g., due to some *external factors*. For instance, consider a scenario where there is only $N = 1$ service type / service queue in the system and the statistical characteristics of the arrival process for the single service type of interest is represented by $\psi \triangleq <\lambda, \sigma, \rho(\cdot)>$ which belongs to a discrete set of outcomes Ψ with a probability mass function $f_\Psi(\cdot)$. Here, for the ease of presentation, we dropped subscript i from the mean, variance, and auto-correlation notations. The expected value of the profit is obtained as

$$E\{\text{Profit}\} = \sum_{\psi \in \Psi} [\text{Profit} \mid \psi] \; f_\Psi(\psi). \tag{3}$$

As an example, suppose $\Psi = \{\psi_1, \psi_2\}, f_\Psi(\psi_1) = \beta$, $f_\Psi(\psi_2) = (1-\beta)$, $\psi_1 = <\lambda_1, \sigma_1, \rho_1(\cdot)>$, and $\psi_2 = <\lambda_2, \sigma_2, \rho_2(\cdot)>$. In this case, the profit maximization problem (3) becomes

$$\begin{aligned} \underset{0 \leq \mu \leq \mu_{max}}{\textbf{Maximize}} \;\; & R(T\lambda_1(1 - q_{\psi_1}(\mu))f_\Psi(\psi_1) + \\ & R(T\lambda_2(1 - q_{\psi_2}(\mu))f_\Psi(\psi_2) - C(\mu), \end{aligned} \tag{4}$$

where $q_{\psi_1}(\mu)$ and $q_{\psi_2}(\mu)$ denote the loss probability in the communications system of interest, when the arrival process carries statistical characteristics ψ_1 and ψ_2, respectively.

III. Loss Probability Model

In this section, we present a mathematical model for loss probability $q(\mu)$ to be used in problems similar to (2) and (4).

First, suppose $\mu \geq \lambda$. In [9], the following loss probability model of a G/D/1 queuing system was proposed for this case:

$$q_I(\mu) = \alpha(\mu) \, e^{-\frac{1}{2} \min\limits_{n \geq 1} \; m_n(\mu)}, \tag{5}$$

where

$$\alpha(\mu) = \frac{1}{\lambda\sqrt{2\pi}\sigma} e^{\frac{(\mu-\lambda)^2}{2\sigma^2}} \int_{\mu}^{\infty} (r-\mu) e^{-\frac{(r-\lambda)^2}{2\sigma^2}} dr \quad (6)$$

and for each integer number $n \geq 1$ we have

$$m_n(\mu) = \frac{(L + n(\mu - \lambda))^2}{n\sigma^2 + 2\sum_{l=1}^{n-1} \rho(l)(n-l)}. \quad (7)$$

Here, L denotes the size of the finite queue.

Theorem 1: The loss probability function $q_I(\mu)$ in (5) is *not* convex over interval $\mu \in [\lambda, \lambda + 0.5\sigma]$; however, it is *convex* and *non-increasing* over interval

$$\mu \in [\lambda + 0.5\,\sigma, +\infty]\,. \quad (8)$$

The proof of Theorem 1 is given in Appendix A.

Next, suppose $\mu \leq \lambda$. This scenario may occur, e.g., in stochastic optimization, where service rate is less than the mean arrival rate under certain random scenarios. In this case, the server would always be busy. Accordingly, out of the total $T\lambda$ service requests that are received within the interval of length T, a total of $T\mu$ service requests are handled, while the rest, i.e., $T\lambda - T\mu$ service requests are dropped. Therefore, the loss probability can be approximated as

$$q_{II}(\mu) = \frac{T\lambda - T\mu}{T\lambda} = \frac{\lambda - \mu}{\lambda}. \quad (9)$$

As a special case, if $\mu \to 0$, then $q_{II}(\mu) \to 1$. Note that $q_{II}(\mu)$ is a linear (thus convex) and decreasing function of μ.

Using the empirical data in [10] with $T = 15$ minutes, the accuracy of the loss probability models in (5) and (9) are assessed in Fig. 1. We can see that $q_I(\mu)$ in (5) is accurate when $\mu \to \infty$ and $q_{II}(\mu)$ in (9) is accurate when $\mu \to 0$. However, both models lose accuracy when μ approaches λ, from the right hand side in case of $q_I(\mu)$, and from the left hand side in case of $q_{II}(\mu)$. Therefore, we propose to *adjust* and *combine* the loss probability models (5) and (9) and obtain the following alternative loss probability model:

$$q(\mu) = \begin{cases} q_I(\mu) & \mu_I \leq \mu \\ q'_{I+}(\mu_I)(\mu - \mu_I) + q_I(\mu_I) & \mu_{II} \leq \mu \leq \mu_I \\ q_{II}(\mu) & \mu < \mu_{II}, \end{cases} \quad (10)$$

where $q'_{I+}(\mu_I)$ denotes the *right derivative* of function $q_I(\mu)$ at $\mu = \mu_I$. The point $\mu_I \geq \lambda + 0.5\sigma$ is chosen in a way that, the right tangent to $q_I(\mu)$ at μ_I intersects $q_{II}(\mu)$ at a point μ_{II} such that $\lambda - \sigma \leq \mu_{II} \leq \lambda$. The proof on the guaranteed existence of parameters μ_I and μ_{II} is omitted for brevity.

Theorem 2: The loss probability function $q(\mu)$ in (10) is convex in service rate for its entire operation range $\mu \geq 0$.

The above theorem directly results from Theorem 1 and the way that the loss probability function $q(\mu)$ is constructed.

To examine the accuracy of the proposed loss probability model, next we generate 100 random time series of length $T = 15$ minutes [11, Section 5.1], based on the statistical characteristics of randomly selected 15 minutes intervals of the data in [10]. Fig. 2 shows the mean absolute error (sorted in ascending order) of the proposed loss probability model in (10) and the one in (5) over the interval $[\lambda, \mu_I]$ for these 100 time series. From Fig. 2, the loss probability model in (10) has a lower mean absolute error than the one in (5).

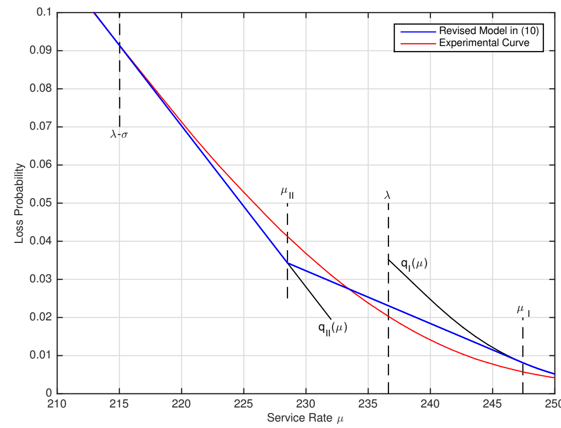

Fig. 1. Loss probability as a function of service rate μ: empirical curve versus the three analytical curves according to (5), (9), and (10).

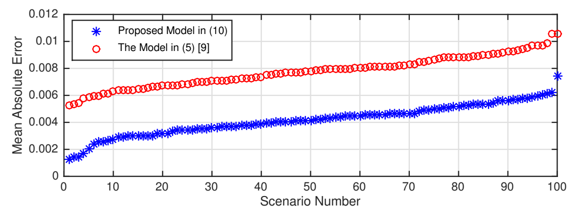

Fig. 2. The mean absolute error of the proposed loss probability model in (10) and that of the one in (5) over 100 randomly generated time series.

Theorem 3: Let us define n_{max} such that $\rho(l) = 0$ for any $l \geq n_{max}$. If $L \geq \sigma n_{max}$, there exist a parameter $\mu^\star \geq \mu_{II}$ such that the proposed loss probability model in (10) is a more accurate approximation of the true loss probability than the model in (5) over the interval $\mu^\star \leq \mu \leq \mu_I$.

The proof of Theorem 3 is given in Appendix B.

IV. Case Studies

First, consider Case 1 in Section II-A. Here, we simulate 30 time slots of length $T = 15$ minutes. We assume that $N = 3$ different types of service requests are handled by the shared server. The service request arrival rates for the first, the second, and the third service types are set based on the World Cup data on June 14th, 15th, and 16th, respectively, from 12:00 AM to 7:30 AM [10]. We set $R_i(x) = 100 w_i log(1+x)$, where $w_1 = 1$, $w_2 = 2$, and $w_3 = 3$ are the *revenue weighting factors*. The cost functions are fixed. Hence, problem (2) reduces to a multi-service queue revenue maximization problem.

Simulation results are shown in Fig. 3, where the operating time is divided into *three* time frames. First, during time slots 1 to 10, there are service requests for service types 1 and 2. After that, during time slots 11 to 20, service requests arrive from all three service types. Finally, during time slots 21 to 30, the server receives requests for service types 1 and 3, but not 2. From the results in Fig. 3, a service rate allocation based on the optimal solution of problem (2) manages the resources based

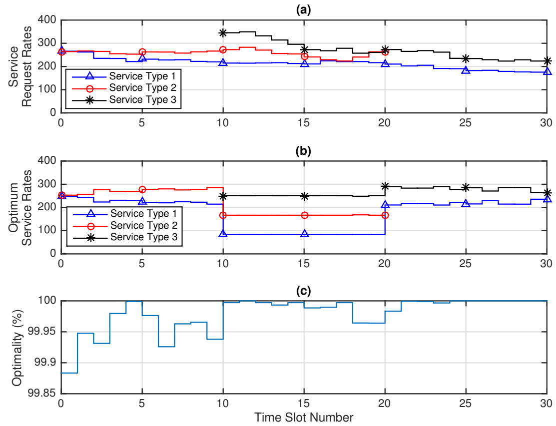

Fig. 3. Simulation results for Case 1: (a) The mean service request arrival rates. (b) The service rates by solving (2). (c) The optimality in comparison with the true optimal profit obtained from an event-based simulation.

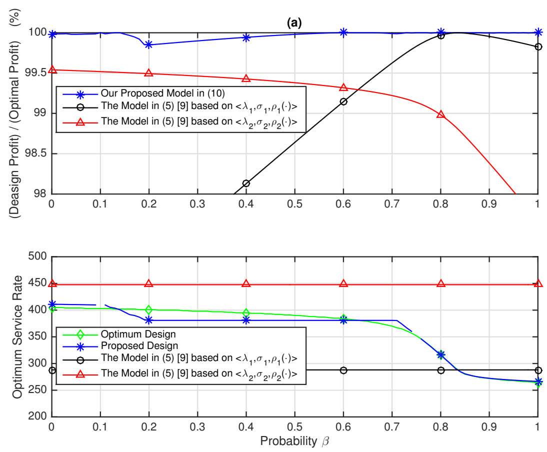

Fig. 4. Simulation results for Case 2: (a) The optimality in maximizing the expected profit. (b) Optimal service rate based on different design approaches.

on the priority of incoming service requests, giving higher service rates to service requests with higher priorities.

Next, consider Case 2 in Section II-B. Think of a web-based video streaming server for a playoff soccer game. Suppose the 90 minutes normal game time is about to finish while the game is in a tie. The server administrators need to allocate resources for the next $T = 15$ minutes. There are *two possibilities*. First, with probability β, one team scores and the game ends in normal time. In that case, the workload to the servers *will drop significantly* as many online viewers do not watch the post-game show. Second, with probability $1-\beta$, the game ends in tie and the *extra time* is implemented. In this case, the workload will remain high as users will continue watching the game. In this example, the outcome of the soccer game is the *external factor* in Section II-B. It is natural to assume that the World Cup server administrators have a good estimate, based on historical data, about the statistical characteristics of the workload during a *post-game show streaming*, i.e., ψ_1, and during a *live game streaming*, i.e., ψ_2. However, they do *not* know which one of these two scenarios will occur. Therefore, they need to solve a stochastic optimization problem as in (4). Here, we use the statistical characteristics of the workload data [10] during time slot 4:15 AM to 4:30 AM on June 29th to obtain ψ_1 and the statistical characteristics of the workload data during time slot 2:00 AM to 2:15 AM on June 24th to obtain ψ_2. Note that, we have $\lambda_1 = 248.48$ and $\lambda_2 = 394.77$.

The *expected* profit for the next $T = 15$ minutes are shown in Fig. 4(a), where the probability parameter β varies from 0 to 1. Here, the expected value is calculated by examining 100 random workloads that are generated based on the statistical characteristics that we obtained from the empirical data, using the *time series generation* scheme in [11, Section 5.1]. We can see that, our proposed stochastic optimization approach can accurately maximize the expected profit for all values of β. In contrast, the model in (5) [9] cannot be used for stochastic optimization. Instead, we should either use $<\lambda_1, \sigma_1, \rho_1(\cdot)>$ or $<\lambda_2, \sigma_2, \rho_2(\cdot)>$; either way, the results will be represented by flat curves as in Fig. 4(b). For certain values of β, e.g., $\beta = 0.9$, the optimal service rate μ^* is greater than λ_1 but less than λ_2. Therefore, it is necessary to use a loss probability model that works for both $\mu \geq \lambda$ and $\mu \leq \lambda$, as in (10).

V. Conclusions

A convex optimization framework was proposed for service rate allocation in finite communications buffers with example applications to maximum profit multi-service scheduling and stochastic service rate allocation. Using empirical data, both deterministic and stochastic case studies were investigated.

Appendix A: Proof of Theorem 1

Let us define $t \triangleq (\mu - \lambda)/\sigma$. We can reformulate (6) as

$$\alpha(\mu) = \frac{\sigma}{\lambda\sqrt{2\pi}} \left(1 - te^{\frac{t^2}{2}} \int_t^\infty e^{-\frac{u^2}{2}} du\right). \tag{11}$$

Once we take the derivative with respect to μ, we have

$$\alpha'(\mu) = \frac{1}{\lambda\sqrt{2\pi}} \left(t - (t^2+1)e^{\frac{t^2}{2}} \int_t^\infty e^{-\frac{u^2}{2}} du\right). \tag{12}$$

Also, since e^{-x} is non-increasing we have $q_I(\mu) = \max_n q_n(\mu)$, where for each $n \geq 1$, we define

$$q_n(\mu) \triangleq \alpha(\mu)\, e^{-\frac{1}{2}m_n(\mu)}. \tag{13}$$

From [8, Section 3.2.3], $q_I(\mu)$ is proven to be a convex function if we can show that for each $n \geq 1$, we have

$$\begin{aligned} q_n{}''(\mu) = e^{-\frac{1}{2}m_n(\mu)} \Big(&\alpha''(\mu) + \alpha(\mu) {m'_n}^2(\mu)/4 \\ &- \alpha'(\mu)m'_n(\mu) - \alpha(\mu)m''_n(\mu)/2\Big) \geq 0. \end{aligned} \tag{14}$$

We show (14) through the following five steps:

Step 1: Since the function under integral in (6) is positive within the range of integral, from (6) we have $\alpha(\mu) \geq 0$.

Step 2: We show that $\alpha''(\mu) \geq 0$ over interval (8). After taking the second derivative of α with respect to μ, we have

$$\alpha''(\mu) = \frac{(t^3+3t)e^{\frac{t^2}{2}}}{\lambda\sqrt{2\pi}\sigma} \left(\frac{t^2+2}{t^3+3t} e^{\frac{-t^2}{2}} - \int_t^\infty e^{-\frac{u^2}{2}} du\right). \tag{15}$$

Next, we note that

$$\frac{d}{dt}\left(\frac{t^2+2}{t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{+\infty}e^{-\frac{u^2}{2}}du\right)=\frac{-6e^{-\frac{t^2}{2}}}{t^2(t^2+3)^2}<0, \quad (16)$$

and

$$\lim_{t\to+\infty}\left(\frac{t^2+2}{t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{+\infty}e^{-\frac{u^2}{2}}du\right)=0. \quad (17)$$

From (16) and (17), we conclude that for each $t>0$, we have

$$\left(\frac{t^2+2}{t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{+\infty}e^{-\frac{u^2}{2}}du\right)\geq 0. \quad (18)$$

Since $t>0$ for the interval in (8), from (18), we have $\alpha''\geq 0$.

Step 3: We show that

$$-2\alpha'(\mu)/\alpha(\mu)\geq 1/(\sigma t). \quad (19)$$

In other words, from (11) and (12), we need to show that

$$-2\frac{\alpha'(\mu)}{\alpha(\mu)}-\frac{1}{\sigma t}=\frac{1}{\sigma}\left(\frac{2e^{\frac{t^2}{2}}\int_t^{\infty}e^{-\frac{u^2}{2}}du}{1-te^{\frac{t^2}{2}}\int_t^{\infty}e^{-\frac{u^2}{2}}du}-2t-\frac{1}{t}\right)\geq 0.$$

After reordering the terms, we can rewrite this inequality as

$$\frac{2t^2+1}{2t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{\infty}e^{-\frac{u^2}{2}}du\leq 0. \quad (20)$$

Next, we note that

$$\frac{d}{dt}\left(\frac{2t^2+1}{2t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{\infty}e^{-\frac{u^2}{2}}du\right)=\frac{(6t^2-3)e^{-\frac{t^2}{2}}}{(2t^2+3)^2\,t^2}. \quad (21)$$

The above derivative is negative (indicating a decreasing function) for any $0<t<\sqrt{2}/2$ and positive (indicating an increasing function) for any $t>\sqrt{2}/2$. Furthermore, we have

$$\lim_{t\to+\infty}\left(\frac{2t^2+1}{2t^3+3t}e^{-\frac{t^2}{2}}-\int_t^{\infty}e^{-\frac{u^2}{2}}du\right)=0. \quad (22)$$

The function on the left hand side in (20) has a zero at $t=0.466$ and a minimum at $t=\sqrt{2}/2$. From these, together with (21) and (22), we conclude that (20) and consequently (19) hold as long as $t\geq 0.466$, e.g., within the interval in (8).

Step 4: We show that, over interval (8), we have

$$m_n''(\mu)/m_n'(\mu)\leq 1/(\sigma t). \quad (23)$$

From (7) and after taking the derivatives over μ, we have:

$$\frac{m_n''(\mu)}{m_n'(\mu)}=\frac{n}{L+n(\mu-\lambda)}\leq\frac{1}{(\mu-\lambda)}=\frac{1}{\sigma t}. \quad (24)$$

Step 5: From (19) and (23), and since $m_n'(\mu)\geq 0$ [9], we have

$$\begin{aligned} &m_n''(\mu)/m_n'(\mu)\leq -2\alpha'(\mu)/\alpha(\mu)\\ &\quad\Rightarrow\quad -\alpha'(\mu)m_n'(\mu)-\alpha(\mu)m_n''(\mu)/2\geq 0.\end{aligned} \quad (25)$$

Also, from Steps 1 and 2, over interval (8) we have

$$\alpha''(\mu)+\alpha(\mu)\,{m_n'}^2(\mu)/4\geq 0. \quad (26)$$

From (25) and (26) and since the exponential function is non-negative, we can conclude (14) and the proof is complete. ■

Appendix B: Proof of Theorem 3

Since $\rho(l)\leq\sigma^2$ for any l, we have $\sum_{l=1}^{n-1}\rho(l)(n-l)\leq\sigma^2 n(n-1)/2$. From this, together with (7), we can show that

$$\frac{{m_n'}^2}{m_n''}=2\,\frac{(L+n(\mu-\lambda))^2}{n\sigma^2+2\sum_{l=1}^{n-1}\rho(l)(n-l)}\geq 2\left(\frac{L}{n\sigma}+t\right)^2. \quad (27)$$

From (27), if $L\geq\sigma n_{max}$ then

$$\alpha(\mu){m_n'}^2/4-\alpha(\mu)m_n''/2\geq 0. \quad (28)$$

Since $\alpha''(\mu)$ and $-\alpha'(\mu)m_n'$ are non-negative, from (14) and (28), we conclude that $q_n(\mu)$ in (13) is a convex function of μ for any $L\geq\sigma n_{max}$. Accordingly, from [8, Section 3.2.3], $q_I(\mu)$ is convex if $L\geq\sigma n_{max}$. Next, let $\tilde{q}(\mu)$ denote the *true* loss probability at service rate μ. Since $q_I(\mu)$ is an *upper bound* for $\tilde{q}(\mu)$ [9], we have $q(\mu_I)=q_I(\mu_I)\geq\tilde{q}(\mu_I)$. Also, since (9) is a *lower bound* for $\tilde{q}(\mu)$ [12], we have $q(\mu_{II})=q_{II}(\mu_{II})\leq\tilde{q}(\mu_{II})$. From these two facts, and since $\tilde{q}(\mu)$ is a continuous function, $\tilde{q}(\mu)$ must intersect with line segment $q(\mu)$ at a point $\mu_{II}\leq\mu^\star\leq\mu_I$. Consequently, since $\tilde{q}(\mu)$ is convex [13, Theorem 3.1], we conclude that

$$q(\mu)\geq\tilde{q}(\mu)\qquad\forall\mu\in[\mu^\star,\mu_I]. \quad (29)$$

Also, from the convexity of $q_I(\mu)$, we have

$$q_I(\mu)\geq q(\mu)\qquad\forall\mu\geq\lambda. \quad (30)$$

From (29) and (30), the proposed loss probability model in (10) is a *tighter upper bound* for $\tilde{q}(\mu)$ than the model in (5) for any service rate $\mu^\star\leq\mu\leq\mu_I$ when $L\geq\sigma n_{max}$. ■

References

[1] F. Papier, *Optimization of Rental Systems: Queuing Loss Theory for the Optimization of Cargo Vehicle Rental Systems*. cologne germany: Klner Wissenschaftsverlag, 2007.

[2] S. Bu, F. R. Yu, Y. Cai, and X. P. Liu, “When the smart grid meets energy-efficient communications: Green wireless cellular networks powered by the smart grid,” *IEEE Transactions on Wireless Communications*, vol. 11, no. 8, pp. 3014–3024, August 2012.

[3] M. Chiang, A. Sutivong, and S. Boyd, “Efficient nonlinear optimization of queuing systems,” in *Proc. of IEEE Globecom*, Taiwan, Nov 2002.

[4] S. J. Macgregor, “Multi-server, finite waiting room, M/G/c/K optimization models,” *Asia-Pacific Journal of Operational Research*, vol. 25, no. 04, pp. 531–561, 2008.

[5] F. R. B. Cruz, “Optimizing the throughput, service rate, and buffer allocation in finite queueing networks,” *Journal of Electronic Notes in Discrete Mathematics*, vol. 35, no. 1, pp. 163–168, December 2009.

[6] R. Nagarajan and D. Towsley, “A note on the convexity of the probability of a full buffer in the M/M/1/K queue,” *IEEE Journal on Selected Areas in Communication*, pp. 92–85, August 1992.

[7] L. Ast, T. Cinkler, G. Fodor, S. Racz, and S. Blaabjerg, “Blocking probability approximations and revenue optimization in multi-rate loss networks,” *Modelling and Simulation of Computer Systems and Networks*, vol. 68, no. 1, pp. 56–65, January 1997.

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

[9] H. S. Kim and N. B. Shroff, “Loss probability calculations and asymptotic analysis for finite buffer multiplexers,” *IEEE/ACM Trans. on Networking*, vol. 9, no. 6, pp. 755 –768, December 2001.

[10] Http://ita.ee.lbl.gov/html/contrib/WorldCup.html.

[11] D. P. Kroese, T. Taimre, and Z. I. Botev, *Handbook of monte carlo methods*. Hoboken, NJ, USA: John Wiley & Sons Inc., 2011.

[12] D. P. Heyman, “Comments on a queueing inequality,” *Management Science*, vol. 26, no. 9, pp. 956–959, Sep 1980.

[13] K. Kumaran, M. Mandjes, and A. Stolyar, “Convexity properties of loss and overflow functions,” *Operations Research Letters*, vol. 31, no. 2, pp. 95–100, March 2003.