

Technical Appendix for:
Sensitivity of Optimal Capacity to Customer Impatience in an
Unobservable M/M/S Queue
(Why You Shouldn't Shout at the DMV)

Mor Armony¹

Erica Plambeck²

Sridhar Seshadri³

Proof of Theorem 1: First consider the decision variable μ . We first prove the result for the single-server case ($S = 1$), and then deal with the general multi-server case. The proof is based on the sample path approach. Specifically, we prove that Y (viewed as a function of μ) satisfies *sample path convexity* (a term that has been introduced by Shaked and Shanthikumar (1988)). Specifically, let $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ be four service rates such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$, and fix λ , $\beta(\cdot)$ and $\eta(\cdot)$. Suppose that there exist Y_1, \dots, Y_4 , which are versions of the original head-count processes (Y_i has service rate of μ_i) that satisfy the following two properties for all $t \geq 0$:

1. $Y_1(t) + Y_4(t) \geq Y_2(t) + Y_3(t)$, a.s.
2. $Y_1(t) \geq \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

Then, according to Shaked and Shanthikumar (1988), Y is said to be stochastically decreasing and convex in the sample path sense (SDCX(sp)). From Theorem 3.6, Proposition 2.11 and Remark 2.8 of Shaked and Shanthikumar (1988) it follows that $Eh(Y(\infty))$ is decreasing and convex in μ , for any increasing and convex function h . In particular, $EY(\infty)$ is decreasing and convex in μ .

To construct the coupled versions Y_1, \dots, Y_4 we wish to come up with appropriate uniformized discrete versions of the original processes. However, for uniformization to work one needs bounded transition rates of the original Markov chain, which is not the case in this paper (we do *not* assume boundedness of the reneging rates $\eta(y)$). To resolve this problem we define for all $M > 0$ a truncated reneging function $\eta_M(y) = \min\{\eta(y), M\}$. Clearly, since $\eta(\cdot)$ is concave, and $\min\{\cdot, M\}$, is non-decreasing and concave, $\eta_M(\cdot)$ is also concave.

¹Stern School of Business, New York University, marmony@stern.nyu.edu

²Graduate School of Business, Stanford University, elp@stanford.edu

³Stern School of Business, New York University, seshadr@stern.nyu.edu

Moreover, for any fixed $M > 0$, $\eta_M(\cdot)$ is bounded. Let Y_1^M, \dots, Y_4^M be uniformized discrete versions of the head-count processes with arrival rate λ , balking probability function $\beta(\cdot)$, service capacity μ_i , $i = 1, \dots, 4$, and reneging rate function $\eta_M(\cdot)$. We will show that for each $M > 0$ if 1. and 2. are satisfied at time $n = 0$ with respect to Y_1^M, \dots, Y_4^M then they hold at time n for all $n \in \mathbb{Z}_+$. It will then follow that $Eh(Y^M(\infty))$ is decreasing and convex in μ . But since $Y^M(\infty)$ weakly converges to $Y(\infty)$ ⁴ it follows from Proposition 2.11 of Shaked and Shanthikumar (1988) that $Eh(Y(\infty))$ is a decreasing and convex function of μ .

We now fix $M > 0$, and establish, by induction, that if 1. and 2. hold at time $n = 0$ for Y_1^M, \dots, Y_4^M , then they hold for all $n = 1, 2, \dots$. For brevity, we omit the superscript M from the subsequent terms. In addition to 1. and 2. we define a third property as follows:

$$\tilde{1}. Y_1(n) + Y_4(n) = Y_2(n) + Y_3(n),$$

that is, property $\tilde{1}$. is property 1. with an equality replacing the inequality. We first establish that if properties $\tilde{1}$. and 2. are satisfied at time n , then properties 1. and 2. hold at time $n + 1$. Let $v = \lambda + \mu_4 + M$. be an upper bound on the total transition rate of the processes Y_1, \dots, Y_4 . For n , such that $\tilde{1}$. and 2. hold, we define the following possible uniformized and coupled transitions:

Arrival + balking: With probability $\frac{\lambda}{v}$ we have a new order arriving into all four systems.

When a new order arrives, it balks system i with probability $\beta(Y_i(n))$. This is done as follows: Let $Y_{(1)}(n) \geq Y_{(2)}(n) \geq Y_{(3)}(n) \geq Y_{(4)}(n)$ be the order statistics for $Y_i(n)$, $i = 1, \dots, 4$. Respectively, refer to system (i) as the systems whose head-count is $Y_{(i)}(n)$. Note that from properties $\tilde{1}$ and 2, it follows that $Y_{(1)}(n) = Y_1(n)$ and $Y_{(4)}(n) = Y_4(n)$.

Now let $\beta_{(i)} = \beta(Y_{(i)}(n))$. From the monotonicity and concavity of $\beta(\cdot)$ it follows that:

$$\mathbf{a.} \quad \beta_{(1)} \geq \beta_{(2)} \geq \beta_{(3)} \geq \beta_{(4)},$$

$$\mathbf{b.} \quad \beta_{(1)} + \beta_{(4)} \leq \beta_{(2)} + \beta_{(3)}.$$

Now, let $U \sim Uniform(0, 1)$. U will determine in which systems the order just arrived will immediately balk according to the following rules:

i. If $U \leq \beta_{(4)}$, then balk in all four systems.

⁴This can be shown by writing down the stationary distributions of the corresponding birth and death processes explicitly, and show that those distributions converge to the limiting one, with unbounded reneging rates.

- ii. Else, if $U \leq \beta_{(2)} + \beta_{(3)} - 1$, then balk from queues (1), (2) and (3).
- iii. Else, if $U \leq \beta_{(3)}$, then balk in queues (3) and (1) only.
- iv. Else, if $U \leq \beta_{(1)}$, then balk in queues (2) and (1) only.
- v. Else, if $U \leq \beta_{(2)} + \beta_{(3)} - \beta_{(4)}$, balk in queue (2) only.

To verify that the balking occurs according to the right probabilities, note that in systems (1), (3) and (4) the balking probabilities are trivially equal to the required probabilities provided that $\beta_{(4)} = \min\{\beta_{(i)}\}$ and $\beta_{(1)} \geq \beta_{(3)}$, both are guaranteed by a. In queue (2), if $\beta_{(2)} + \beta_{(3)} - \beta_{(4)} < 1$ balking will occur with probability: $\beta_{(4)} + (\beta_{(2)} + \beta_{(3)} - \beta_{(4)} - \beta_{(3)}) = \beta_{(2)}$, provided that $\beta_{(2)} + \beta_{(3)} - \beta_{(4)} \geq \beta_{(1)}$, which is equivalent to b. Similarly, if $\beta_{(2)} + \beta_{(3)} - \beta_{(4)} \geq 1$, balking in this queue will occur with probability: $(\beta_{(2)} + \beta_{(3)} - 1) + (1 - \beta_{(3)}) = \beta_{(2)}$.

Service Completion: With probability $\frac{\mu_4}{v}$ we have a service completion event. To determine which systems are going to indeed have service completions (as opposed to a transition from a state to itself), let $U \sim Uniform(0, 1)$.

- a. If $U < \frac{\mu_1}{\mu_4}$ we have service completions from all systems for which $Y_i(n) > 0$.
- b. If $\frac{\mu_1}{\mu_4} \leq U < \frac{\mu_2}{\mu_4}$, we have departures in systems 2 and 4 only, whenever the corresponding queues are non-empty.
- c. If $\frac{\mu_2}{\mu_4} \leq U < 1$, we have departures in systems 3 and 4 only, whenever the corresponding queues are non-empty.

It is easy to see, that system i has a service completion with probability $\frac{\mu_i}{v}$ as long as $Y_i(n) > 0$ (recall that $\mu_1 + \mu_4 = \mu_2 + \mu_3$). Note that the reason why we do not simply have a service completion from system i whenever $U < \frac{\mu_i}{\mu_4}$, is that in this case we may have a service completion from system 4 only, which may violate property 1.

A Reneging Job (order cancellation) : Finally, with probability $[\eta(Y_i(n)) \wedge M]/v$ we have an order cancellation from system i . The coupling works as follows: let $Y_{(1)}(n) \geq Y_{(2)}(n) \geq Y_{(3)}(n) \geq Y_{(4)}(n)$ be the ordered statistics of $Y_1(n), \dots, Y_4(n)$, and let $\xi_{(i)} = \eta_M(Y_{(i)}(n)) = \min\{\eta(Y_{(i)}(n)), M\}$. Note that property $\tilde{1}$. and the convexity of $\eta_M(\cdot)$ imply that $\xi_{(1)} + \xi_{(4)} \leq \xi_{(2)} + \xi_{(3)}$ (that is, the inequality with respect to the ξ_i 's is the opposite of property 1.) Let $U \sim Uniform(0, 1)$ be the random variable that determines the reneging from all systems. Let $m = \max\{M, \xi_{(3)} + \xi_{(2)} - \xi_{(4)}\}$.

- a. If $U < \frac{\xi_{(4)}}{m}$, we will have one order cancellation from all the systems such that $Y_i(n) > 0$.
- b. If $\frac{\xi_{(4)}}{m} \leq U < \frac{\xi_{(3)}}{m}$, we have one order cancellation from each of the systems (3) and (1) (provided that $Y_{(i)}(n) > 0$, for $i = 1, 3$).
- c. If $\frac{\xi_{(3)}}{m} \leq U < \frac{\xi_{(1)}}{m}$, we have one order cancellation from each of the systems (2) and (1) (provided that $Y_{(i)}(n) > 0$, for $i = 1, 2$).
- d. If $\frac{\xi_{(1)}}{m} \leq U < \frac{\xi_{(3)} + \xi_{(2)} - \xi_{(4)}}{m}$, we have one order cancellation from system (2), provided that $Y_{(2)}(n) > 0$.

Note that given this setup, an order cancellation occurs in system (i) with probability $[\eta(Y_{(i)}(n)) \wedge M]/v$.

We will now show that if properties $\tilde{1}$. and 2. hold at time n , 1.-2. are satisfied at time $n + 1$. We will go over the different types of events, to show that 1.-2. still hold at time $n + 1$:

Arrival + balking: Since we have arrivals coming into all systems at the same time, properties 1.-2. will still hold at time $n + 1$, if no balking occurs. To verify that properties 1. and 2. hold at time $n + 1$ in case of balking note that those properties might be violated only if from time n to $n + 1$ one of the following occurs:

- I. The LHS of 1. stays the same, while the RHS of 1. increases by 1 or 2: This will only occur when there is balking in both queues (1) and (4), which implies balking in queues (2) and (3) as well.
- II. The LHS of 1. increases by 1, while the RHS of 1. increases by 2: This change in the LHS of 1. can only occur when the arrival to queue (4) does not balk, while the arrival to queue 1 balks. However, in this case, at least one of the arrival to queue (2) or (3) will balk.
- III. $Y_i(n) = Y_1(n)$ for some $i \neq 1$, and Y_1 stays the same, while Y_i increases by 1 (this will violate 2.): This would occur only if $Y_{(2)}(n) = Y_{(1)}(n)$ and there will be balking in queue (1) and not in queue (2) (recall that queue (1) and queue 1 are the same). However, if $Y_{(2)}(n) = Y_{(1)}(n)$, then $Y_{(3)}(n) = Y_{(4)}(n)$, and in particular $\beta_{(3)} = \beta_{(4)}$ and $\beta_{(2)} = \beta_{(1)}$. In this case, it is easily verified that balking in queue (1) implies balking in queue (2) as well.

Service Completion: Here we have to make sure we are avoiding the following:

- I. The LHS of 1. decreases by 1, while the RHS does not change:
- II. The LHS of 1. decreases by 2, while the RHS decreases by 1 or does not change.
- III. $Y_i = Y_1$ for some $i \neq 1$, and Y_1 decreases by 1, while Y_i does not change (hence property 2. is violated).

Observe that none of these can happen because whenever $Y_i = 0$ for either $i = 2$ or 3 , we have $Y_4 = 0$. Moreover, if $Y_i = Y_1$, then if Y_1 decreases, Y_i will also decrease.

Order Cancellation: In this case, properties 1. - 2. will be violated if any of the above I.-III. occur. We show that this cannot happen by going over the different values of the uniform variable U . First note that here $Y_{(1)} = Y_1$ and $Y_{(4)} = Y_4$. Without loss of generality, assume that $Y_{(2)} = Y_2$, and $Y_{(3)} = Y_3$, and omit the (\cdot) from the subscript. Also, recall that $\xi_1 + \xi_4 \leq \xi_2 + \xi_3$.

- a. If $U < \frac{\xi_4}{m}$, then $Y_4(n) > 0$, which implies that $Y_i(n) > 0$ for all i , which means that all values of $Y_i(n)$ will be reduced by 1.
- b. If $\frac{\xi_4}{m} \leq U < \frac{\xi_3}{m}$, then $Y_3(n) > Y_4(n)$. This implies that $Y_1(n) > Y_2(n)$ (from property $\tilde{1}$.), and therefore the fact that $Y_1(n)$ and $Y_3(n)$ are the only processes reduced by 1, will not violate 1.-2.
- c. If $\frac{\xi_3}{m} \leq U < \frac{\xi_1}{m}$, then $Y_1(n) > Y_3(n)$. This implies that $Y_2(n) > Y_4(n) \geq 0$ (see property $\tilde{1}$.), and therefore the fact that $Y_1(n)$ and $Y_2(n)$ are the only processes reduced by 1, will not violate 1.-2.
- d. If $\frac{\xi_1}{m} \leq U < \frac{\xi_2 + \xi_3 - \xi_4}{m}$, then 1. - 2. will clearly not be violated.

So far we have shown that if at time n properties $\tilde{1}$. and 2. hold, then at time $n + 1$ both properties 1 and 2 will hold. Suppose that at time n property 1. holds with a strict inequality, that is:

$$Y_1(n) + Y_4(n) > Y_2(n) + Y_3(n).$$

In order to describe the transitions in this case, we first define the following transformation of $Y_1(n)$ and $Y_4(n)$: $\tilde{Y}_4(n) = \min\{Y_i(n), i = 1, 2, 3, 4\}$ and $\tilde{Y}_1(n) = Y_1(n) - (Y_1(n) + \tilde{Y}_4(n) - (Y_2(n) + Y_3(n)))$. For convenience of notation also let $\tilde{Y}_2(n) = Y_2(n)$ and

$\tilde{Y}_3(n) = Y_3(n)$. It is easy to see that $\tilde{Y}_i(n) \leq Y_i(n)$ for $i = 1, 4$, that $\tilde{Y}_1(n) = \max\{\tilde{Y}_i(n), i = 1, 2, 3, 4\}$ and that $\tilde{Y}_1(n) + \tilde{Y}_4(n) = \tilde{Y}_2(n) + \tilde{Y}_3(n)$. That is, properties $\tilde{1}$. and 2. hold for the modified values of $Y_i(n)$. Let $\tilde{Y}_i(n+1), i = 1, 2, 3, 4$, be the values of these processes after one transition, that occurred according to the above rules. In particular, we know that properties 1. and 2. hold for $\tilde{Y}_i(n+1), i = 1, 2, 3, 4$. Let $\bar{F}_{i,x}(y) = P_{\mu=\mu_i}\{Y_i(n+1) > y \mid Y_i(n) = x\}$, then it is easy to verify that $\bar{F}_{i,x}(y)$ is non-decreasing in x . For a cdf \bar{F} define \bar{F}^{-1} , the inverse of \bar{F} , as $\bar{F}^{-1}(u) = \inf\{x \mid \bar{F}(x) \leq u\}$, $0 \leq u < 1$. In particular, for $i = 1, 4$, let $Y_i(n+1) = \bar{F}_{i,Y_i(n)}^{-1}(\bar{F}_{i,\tilde{Y}_i(n)}(\tilde{Y}_i(n+1)))$, and for $i = 2, 3$, simply let $Y_i(n+1) = \tilde{Y}_i(n+1)$. One can now easily verify that for all i , $Y_i(n+1) \geq \tilde{Y}_i(n+1)$, that properties 1. and 2. hold for $Y_i(n+1), i = 1, \dots, 4$, and that $Y_i(n+1)$ has the right distribution (i.e. for all y , $P_{\mu=\mu_i}\{Y_i(n+1) > y \mid Y_i(n)\} = \bar{F}_{i,Y_i(n)}(y)$). This completes the proof of the theorem for the single server case.

It is left to prove the theorem for the general multiserver ($S > 1$) case. To extend the above proof to the M/M/S system, the only case that needs to be treated is service completions. Let $Y_1 + Y_4 = Y_2 + Y_3$, $Y_1 \geq \max\{Y_2, Y_3, Y_4\}$, $\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ and $\mu_1 + \mu_4 = \mu_2 + \mu_3$. Then, we shall show that

$$\mu_1 \min\{Y_1, S\} + \mu_4 \min\{Y_4, S\} \leq \mu_2 \min\{Y_2, S\} + \mu_3 \min\{Y_3, S\}. \quad (\text{A1})$$

To see this consider the following cases:

- 1) If all Y 's are greater than or equal to S , then (A1) is true by assumption.
- 2) If only $Y_4 < S$, then, again, (A1) is true by assumption.
- 3) If only Y_4 and Y_3 (wlog $Y_2 \geq Y_3$) are less than S then if (A1) is false then

$$\mu_1 S + \mu_4 Y_4 > \mu_2 S + \mu_3 Y_3. \quad (\text{A2})$$

But

$$\mu_1 S + \mu_4 S \leq \mu_2 S + \mu_3 S, \quad (\text{A3})$$

and (A2)-(A3) yields $\mu_4(S - Y_4) < \mu_3(S - Y_3)$, which is false because $Y_4 \leq Y_3$ and $\mu_4 \geq \mu_3$.

- 4) If only Y_2, Y_3, Y_4 are less than S then if (A1) is false then

$$\mu_1 S + \mu_4 Y_4 > \mu_2 Y_2 + \mu_3 Y_3. \quad (\text{A4})$$

However, (recall wlog $Y_2 \geq Y_3$) $\mu_1 Y_1 + \mu_4 Y_4 \leq \mu_1 Y_2 + \mu_4 Y_3$ (because we may increase Y_4 by equal amount as the decrease in Y_1) $\leq \mu_2 Y_2 + \mu_3 Y_3$ (because we may increase μ_1 by same amount as the decrease in μ_4). Thus,

$$\mu_1 Y_1 + \mu_4 Y_4 \leq \mu_2 Y_2 + \mu_3 Y_3. \quad (\text{A5})$$

Then, (A4) and (A5) lead to a contradiction.

5) If all Y 's are less than S , then (A1) follows from (A5).

Finally, (A1) implies that we can couple the four systems such that the second and third have more service completions on each sample path and that $\tilde{1}$. and 2. hold at each service completion.

Now consider the number of servers. Fix μ and suppose that $0 < S_1 \leq S_2 \leq S_3 \leq S_4$, and $S_1 + S_4 = S_2 + S_3$. We show that there exist versions of the corresponding head-count processes such that for all $t \geq 0$

1. $Y_1(t) + Y_4(t) \geq Y_2(t) + Y_3(t)$. a.s., and
2. $Y_1(t) \geq \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

The proof is identical to that for μ as the decision variable, except for the service completion step. Here, too, we have to verify that if 1. or 2. hold with equality at time n , both of them are still true at time $n + 1$. Let $Z_i = S_i \wedge Y_i$, then Z_i is the number of busy servers in system i , $i = 1, 2, 3, 4$. We claim that if $Y_1(n) + Y_4(n) = Y_2(n) + Y_3(n)$ then

$$Z_1(n) + Z_4(n) \leq Z_2(n) + Z_3(n), \quad (\text{A6})$$

and that if $Y_1(n) = Y_i(n)$ for some $i = 2, 3$ or 4 , then

$$Z_1(n) \leq Z_i(n). \quad (\text{A7})$$

Both (A6) and (A7) can be easily verified by examining all the different possible combinations of $Z_i = S_i$ or $Z_i = Y_i$, for $i = 1, 2, 3, 4$. Now that we have (A6) and (A7) we can couple service completions in a way that 1. and 2. will not be violated as follows: Suppose that 1. holds with an equality. We claim that either $Z_1 = \min\{Z_i\}$ or $Z_4 = \min\{Z_i\}$, $i = 1, 2, 3, 4$. To see that, suppose that $Z_2 = \min\{Z_i\}$. This implies that $Y_2 \leq S_2$, because $S_2 \geq S_1 \geq Z_1$. But in this case, $Y_4 \leq Y_2 \leq S_2 \leq S_4$. In particular, $Z_4 = Y_4 = \min\{Z_i\}$. Similarly, we can argue that if $Z_3 = \min\{Z_i\}$, then $Z_4 = Z_3 = \min\{Z_i\}$. Suppose that $Z_1 = \min\{Z_i\}$, then

1. with probability $Z_1/(Z_2 + Z_3)$ we have departures out of all systems.
2. with probability $(\min\{Z_2, Z_4\} - Z_1)/(Z_2 + Z_3)$ we have departures from systems 4 and 2 only,
3. with probability $(Z_4 - Z_2)^+/(Z_2 + Z_3)$ we have departures from systems 4 and 3 only.
4. With probability $(Z_2 - Z_4)^+/(Z_2 + Z_3)$ we have a departure from system 2 only.
5. Finally, with probability $(Z_3 - Z_1 - (Z_4 - Z_2)^+)/ (Z_2 + Z_3)$ we have a departure from system 3 only.
6. Notice, that with probability $Z_1/(Z_2 + Z_3)$ no transition occurs.

If $Z_4 = \min\{Z_i\}$ then the transitions parallel the ones above, switching the roles of systems 1 and 4.

Two important observations with respect to the above transitions are that (a) All departures from systems 1 or 4 are coupled with departures from systems 2 or 3, and (b) If $Y_i = Y_1$ for some $i = 2, 3$, or 4, then any departure from system 1 will be coupled with a departure from system i . This is true from the description of the transitions above if $Z_1 = \min\{Z_i\}$. In the case that $Z_4 = \min\{Z_i\}$, due to (A7), still any departure from system 1 will be coupled with a departure from system i . Specifically, if, for example, $Y_1 = Y_2$, then by (A7) $Z_1(n) \leq Z_2(n)$. The case in which we might have a departure out of system 1 which is not coupled with a departure from system 2 is case 3. But that happens with probability $(Z_1 - Z_2)^+/(Z_2 + Z_3)$ which is equal to zero. On the other hand, if $Y_1 = Y_3$, then the transition in case 2 would be from systems 1 and 2 but not 3. This happens with probability $(\min\{Z_2, Z_1\} - Z_4)/(Z_2 + Z_3)$. But if $Y_1 = Y_3$ then $Y_2 = Y_4$. In this case, $Z_4 = \min\{Z_i\} \leq Z_2 = \min\{Y_2, S_2\} \leq \min\{Y_4, S_4\} = Z_4$. In particular, $Z_2 = Z_4 \leq Z_1$, so the probability $(\min\{Z_2, Z_1\} - Z_4)/(Z_2 + Z_3)$ is equal to 0. Therefore, the induction step is confirmed under service completion transitions, which completes the proof of the theorem. \square

Proof of Proposition 1: Fix S , λ and η and let $f(\mu) = \eta EY(\infty)$. Clearly, the convexity of f in μ implies that C is also convex in μ . To establish the convexity of f , note that the reneging rate is non-decreasing and concave, and therefore, by Theorem 1 it follows that $f(\mu)$ is convex in μ . The proof of convexity in S is analogous. \square

Proof of Proposition 2: Suppose that $\mu \geq \eta$. By Theorem 3 the head-count process Y is stochastically decreasing and convex in μ and in S . This implies that $Y - S$ is also stochastically decreasing and convex in those two variables. Since the function $f(y - S) = \eta(y - S)^+$ is increasing and convex in $y - S$, it follows that the process $\eta(Y - S)^+$ is also stochastically decreasing and convex in μ and S . Finally, it follows that $\eta E[Y(\infty) - S]^+$ is decreasing and convex in μ . The proof of the proposition is now immediate. \square

Details of Remark 1: In the model studied by Koole and Pot (2006) there are two integer valued decision variables. Their objective function is similar to ours except it has no balking. Also, their proof methodology is different. The two decision variables are s , the number of servers and n the maximal queue size. Let $s(n)$ and $n(s)$ be the optimal values of each of these variables, when the other is fixed. They show that if s were fixed the objective function is increasing in n up to its optimal value. They also show that the optimal n is increasing in s . Finally, they show that the objective function is not convex in the number of servers. To see why this is the case, note that when the decision variables s and n are integer valued and even if the function is jointly integer convex in s and n , the objective function in only one variable might be non-convex. Their counter example demonstrates that for a fixed s we might be forced to use too large a value of n (due to integrality constraint) such that for that large value of n it is better to have a higher value of s . This can happen with any convex function. Note, though, that the absence of convexity in s does not imply that our method is not applicable to their model. In fact, for fixed s or n we believe our method would apply.

Proof of Proposition 3: This proof follows the sample path approach. In particular, we discretize time, and uniformize the transition rates in an analogous way to what was done in the proof of Theorem 1. Specifically, we bound the reneging rate from above by M , and after we prove the result for any M , we let $M \rightarrow \infty$, to get the desired result. Given a value of M , we show that if sample-path sub-modularity holds at time $n = 0$ then it holds for all n . More specifically, suppose that the following three properties hold at time $n = 0$, for all $\lambda_L < \lambda_H$ and $\mu_L < \mu_H$:

I. $Y_{\lambda_H, \mu_L}(n) = \max\{Y_{\lambda, \mu}(n) ; \lambda = \lambda_L, \lambda_H, \mu = \mu_L, \mu_H\},$

II. $Y_{\lambda_L, \mu_H}(n) = \min\{Y_{\lambda, \mu}(n) ; \lambda = \lambda_L, \lambda_H, \mu = \mu_L, \mu_H\},$

$$\text{III. } Y_{\lambda_H, \mu_H}(n) - Y_{\lambda_H, \mu_L}(n) \leq Y_{\lambda_L, \mu_H}(n) - Y_{\lambda_L, \mu_L}(n),$$

then we show by induction that they hold for all $n \geq 0$.

Suppose that $S = 1$, and let $v = \lambda_H + \mu_H + M$. That is, v is the maximal transition rate in all four systems given any state. Now suppose that I.-III. hold at time n , where III. holds with an equality (we will call this property $\widetilde{\text{III}}$). In this case we have three types of transitions:

Arrival: With probability $\frac{\lambda_H}{v}$ we have an arrival event. The coupling works as follows: let $U \sim \text{Uniform}(0, 1)$.

1. If $U < \frac{\lambda_L}{\lambda_H}$ we have one arrival into each of the four systems.
2. If $U \geq \frac{\lambda_L}{\lambda_H}$ we have arrivals into the systems with $\lambda = \lambda_H$ only.

Service Completion: With probability $\frac{\mu_H}{v}$ we have a service completion event. To determine which systems have a departure, let $U \sim \text{Uniform}(0, 1)$.

1. If $U < \frac{\mu_L}{\mu_H}$ we have a service completion for each one of the systems for which $Y_{\lambda, \mu}(n) > 0$.
2. If $U \geq \frac{\mu_L}{\mu_H}$ we have a service completion for those systems with $\mu = \mu_H$ only, whenever $Y_{\lambda, \mu_H}(n) > 0$.

Order Cancellation: With probability $\frac{M}{v}$ we have an order cancellation event. Let $\eta_M(y) = \min\{\eta y, M\}$ be the reneging rate function. Let $Y_{(i)}, i = 1, 2, 3, 4$ be a permutation of $\{Y_{\lambda, \mu}(n); \lambda = \lambda_L, \lambda_H, \mu = \mu_L, \mu_H\}$ such that $Y_{(1)} \geq Y_{(2)} \geq Y_{(3)} \geq Y_{(4)}$. Let $\xi_{(i)} = \eta_M(Y_{(i)})$. Note that I., II, and $\widetilde{\text{III}}$. and the concavity of $\eta_M(\cdot)$ imply that $\xi_{(1)} + \xi_{(4)} \leq \xi_{(2)} + \xi_{(3)}$. Finally, let $m = \max\{M, \xi_{(2)} + \xi_{(3)} - \xi_{(4)}\}$. To determine which systems have a service cancellation, let $U \sim \text{Uniform}(0, 1)$.

1. If $U < \frac{\xi_{(4)}}{m}$, we have a service cancellation from each one of the systems, provided that the corresponding head-count is positive.
2. If $\frac{\xi_{(4)}}{m} \leq U < \frac{\xi_{(3)}}{m}$, we have a service cancellation in systems (1) and (3), provided that $Y_{(i)} > 0, i = 1, 3$.
3. If $\frac{\xi_{(3)}}{m} \leq U < \frac{\xi_{(1)}}{m}$, we have a service cancellation in systems (1) and (2), provided that $Y_{(i)} > 0, i = 1, 2$.

4. If $\frac{\xi_{(1)}}{m} \leq U < \frac{\xi_{(2)} + \xi_{(3)} - \xi_{(4)}}{m}$, we have a service cancellation in system (2), provided that $Y_{(2)} > 0$.

Verifying that if $I.$, $II.$, and \widetilde{III} hold at time n , then $I.$, $II.$ and $III.$ hold at time $n + 1$ is straightforward, and is analogous to proving Theorem 1. We omit the details. If instead of \widetilde{III} , we have III at time n , proceed similarly to the proof of the Theorem 1 to validate the induction step. If $S > 1$ proceed similarly to the general proof of Theorem 1 realizing that the only case to be concerned about is the service completion. However, the service rates of the four systems being compared can be ordered as μ_1, \dots, μ_4 as in the proof of Theorem 1. Therefore, this part of the proof extends without modifications.

This completes the proof of the proposition. \square

Proof of Proposition 4: Consider four systems. The arrival rate and the number of servers in these are: (1) (λ_H, S_L) , (2) (λ_L, S_L) , (3) (λ_H, S_H) , and (4) (λ_L, S_H) . The service rate, and reneging rate in all four systems are the same. Uniformize and couple the transitions in the four systems. Let $Y_i(n)$ be the number of customers in system i at the end of the n th event in the coupled systems. Similarly to the proof of Proposition 3 we wish to establish that:

I. $Y_1(n) = \max\{Y_i(n) ; i = 1, 2, 3, 4\},$

II. $Y_4(n) = \min\{Y_i(n) ; i = 1, 2, 3, 4\},$

III. $Y_1(n) + Y_4(n) \geq Y_2(n) + Y_3(n).$

Note that we only need to consider service completion events because the dynamics of other events are analogous to the ones analyzed in the proof of Proposition 3. Thus, the only event considered below is service completion.

We claim that, without loss of generality, III. can be assumed to hold as an equality. This is true because if $Y_1(n) + Y_4(n) > Y_2(n) + Y_3(n)$, for some n (we henceforth omit the index n from the notation), then let $\tilde{Y}_1 = Y_1 - (Y_1 + Y_4 - (Y_2 + Y_3))$. Then the four systems with \tilde{Y}_1 in place of Y_1 satisfies I. II. and III., where the latter is satisfied with equality. Thus, we couple the four original systems pretending that $Y_1 = \tilde{Y}_1$, while the potential departures of the remainder $Y_1 - \tilde{Y}_1$ customers in system 1 are decoupled from departures in other systems. Hence, we now assume, without loss of generality, that at time n properties I. and II. hold along with the property:

$$\widetilde{\text{III}}. Y_1(n) + Y_4(n) = Y_2(n) + Y_3(n).$$

Consider first the case when all servers are busy in all four systems. In this case, we can simply couple the transitions in the first S_L servers in all systems. Thereafter, couple the transitions in the $S_H - S_L$ remaining servers in each of systems 3 and 4 with one another. Thus, Y_1 will decrease only when all other Y_i 's decrease. Also, Y_4 will decrease whenever any of the other Y_i 's decrease. Thus, Y_4 stays the smallest. Y_1 stays the largest. Also, the difference between $Y_4 + Y_1 - (Y_2 + Y_3)$ is preserved.

Consider the case when $Y_i \leq S_i$ (where S_i denotes the number of servers in system i) in all four systems. Couple the transitions of the first Y_4 servers in all four systems (can be done as Y_4 is the smallest). Thereafter, alternately couple the transitions of a server from systems 2 and 3 with that of a server from system 1. Notice upon a service completion that Y_4 remains the smallest. If, before the transition $Y_1 = Y_2$ (or $Y_1 = Y_3$), then necessarily $Y_4 = Y_3$ (or $Y_4 = Y_2$, respectively) and therefore any transition in system 1 will be coupled with a transition from system 2 (or 3, respectively). Therefore, after the transition Y_1 remains the largest. Also, because the number of transitions (either 1 or 2) in the coupled servers in systems 1 and 4 equal those in systems 2 and 3, $Y_1 + Y_4 - (Y_2 + Y_3)$ remains the same due to transitions in the coupled servers.

Now, consider the case that some $Y_i > S_i$ and some are not. We quickly realize that this can only happen if $Y_1 \geq S_1$. Note that $S_1 + S_4 = S_2 + S_3$ and hence, whenever $\widetilde{\text{III}}$ is satisfied we have that the minimum number of busy servers is attained either in system 4 in system 1. (see the discussion following (A6) and (A7) in the proof of Theorem 1.)

In the former case, couple the Y_4 servers in all systems, then alternately couple a server in system 2 and 3 with one in system 1. Due to (A6) it follows that there might be some additional busy servers in systems 2 and 3. These service completions are decoupled from the other systems. Note that if $Y_1 = Y_2$ (or $Y_1 = Y_3$) then necessarily $Y_4 = Y_3$ (or $Y_4 = Y_2$, respectively). Therefore, by (A7) any transition in system 1 can be coupled with a transition in system 2 (or 3, respectively) and Y_1 will remain the largest. It can be verified that the other relations between the Y_i 's are also preserved.

In the latter case, the number of busy servers is the smallest in system 1. Thus, the number of busy servers in the system 2 is also S_L . Also, because Y_4 is the smallest and due to the assumption that some servers are idle in at least one system, it must be true that $Y_4 < S_H$ whereas, $Y_3 \geq Y_4$. The coupling is straightforward now: couple the busy S_L servers

in all four systems. Couple any remaining busy servers in system 4 with the remaining busy servers in system 3. If there are more busy servers in system 3, they work independently of the rest. Clearly, the service completions preserve the desired inequality, particularly, $Y_2 + Y_3$ can only decrease faster than $Y_1 + Y_4$. Y_1 decreases if and only if a similar decrease takes place in the rest of the Y_i 's and Y_3 (and certainly Y_2) cannot become smaller than Y_4 due to a service completion. \square

Proof of Theorem 3: First consider the parameter region $\mu \geq \eta$ and the case of μ as decision variable. Following the notation of the proof of Theorem 1, let $0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$ be four service rates such that $\mu_1 + \mu_4 = \mu_2 + \mu_3$. Assume that the reneging rate η is bounded above by μ_4 . This is a weaker condition than the condition $\mu \leq \eta$ stated in the Theorem, but it turns out to be sufficient in establishing its results.

Analogously to the proof of Theorem 1, let Y_1, \dots, Y_4 be discretized and uniformized versions of the head-count with service rates μ_1, \dots, μ_4 , respectively, that satisfy properties: **1.** $Y_1(n) + Y_4(n) \geq Y_2(n) + Y_3(n)$, a.s. and **2.** $Y_1(n) \geq \max\{Y_2(n), Y_3(n), Y_4(n)\}$, a.s. at time $n = 0$. By induction, we wish to show that properties 1. and 2. hold for all $n \geq 0$.

The induction proof of 1. and 2. goes through by the simple construction explained next. Note that arrivals and service completions do not introduce a problem. For reneging, we use a slightly different method involving a transfer of a customer. The transfer need not occur physically, but it helps to visualize why we need the condition that $\eta \leq \mu_4$. (We could use the same proof used earlier without this device.) Specifically, transferring a customer from system 1 to system 4, is equivalent to comparing the current set of systems with another set that has $Y_1 - 1$ and $Y_4 + 1$ customers, which in turn is comparable to systems 2 and 3. The idea is that because Y_4 is smaller than Y_1 , if we transfer a customer from system 1 to 4, it will change the total departure rate in both systems combined in the desired direction (either through service completions or through reneging). For example, if both systems have idle servers then the combined rate of departures due to service will increase. If only system 4 has an idle server then the rate will increase because the reneging rate is smaller than μ_4 . Finally, if both are busy, then the total reneging rate remains the same.

The details of the customer transfer are as follows: one can transfer customers from system 1 to system 4 until one of two events happens: either Y_4 equals the minimum of Y_2 and Y_3 , or Y_4 equals S . In the first case, after the transfer Y_1 will equal the maximum of Y_2 and Y_3 . In the second case all systems will have S or more customers. The transfer will not decrease the rate at which queues deplete in systems 1 and 4 due to the assumption on the reneging

rate. Moreover, $\tilde{1}$. (or 1.) and 2. will continue to hold. It thus follows that the induction proof goes through after this modification. In detail, in the first case the two sets of systems will have equal reneging rate. In the second case, $(Y_1 - S) + (Y_4 - S) = (Y_2 - S) + (Y_3 - S)$. The reneging rates depend on these four quantities and the earlier proof for Theorem 1 goes through.

With respect to μ as a decision variable, it is left to establish that Y is stochastically decreasing in μ for $\mu \leq \eta$. This proof is a trivial application of the sample-path coupling approach. Details are omitted. The detailed verification of the induction step for S as a decision variable is as follows:

- (i) Suppose $\eta \leq \mu$. We show that if $S_1 \leq S_2 \leq S_3 \leq S_4$ with $S_1 + S_4 = S_2 + S_3$, then there exist versions of the head-count processes: Y_1, \dots, Y_4 , respectively, such that for all $t \geq 0$
1. $Y_1(t) + Y_4(t) \geq Y_2(t) + Y_3(t)$. a.s., and
 2. $Y_1(t) \geq \max\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

The proof follows similar steps to the proof of Theorem 1. Arrivals to all systems at the same time will not change properties 1. and 2. We now examine departure events, where those include both reneging and service completion. Let $R_i = \eta Q_i + \mu Z_i$, where $Q_i = [Y_i - S_i]^+$ is the queue length in system i , and $Z_i = Y_i \wedge S_i$ is the number of busy servers in the same system. Then, R_i is the total departure rate from system i . Suppose that at time n ,

$$Y_1(n) + Y_4(n) = Y_2(n) + Y_3(n). \quad (\text{A8})$$

We claim that this implies that

$$R_1(n) + R_4(n) \leq R_2(n) + R_3(n). \quad (\text{A9})$$

It is easy to establish (A9) using (A8), (9), and the assumption that $\eta \leq \mu$. We also claim that if $Y_1(n) = Y_i(n)$ for some $i = 2, 3$ or 4 , then $R_1 \leq R_i$ (we omit n from the notation for brevity). This easily follows from (10) and from the assumption that $\eta \leq \mu$. We now define the departure transition. Note that when we say that a departure occurs from system i , it can be either an abandonment (w.p. $\eta Q_i / R_i$) or a service completion (w.p. $\mu Z_i / R_i$). Suppose that $R_2 = \min\{R_i\}$. In this case, $R_3 = \max\{R_i\}$ and the departures are as follows:

- a. With probability $\frac{R_2}{R_2+R_3}$ we have departures out of all systems.
- b. With probability $\frac{R_1-R_2}{R_2+R_3}$ we have departures out of systems 1 and 3 only.
- c. With probability $\frac{R_4-R_2}{R_2+R_3}$ the departures are out of systems 4 and 3 only.
- d. With probability $\frac{R_3+R_2-(R_1+R_4)}{R_2+R_3}$ only a departure out of system 3 occurs.
- e. With probability $\frac{R_2}{R_2+R_3}$ no transitions occur.

Similar transitions are defined in the case that $R_3 = \min\{R_i\}$, switching the roles of systems 2 and 3.

If $R_1 = \min\{R_i\}$ then the transitions are defined as follows:

- a. With probability $\frac{R_1}{R_2+R_3}$ we have departures out of all systems.
- b. With probability $\frac{\min\{R_2, R_4\}-R_1}{R_2+R_3}$ we have departures out of systems 2 and 4 only.
- c. With probability $\frac{[R_4-R_2]^+}{R_2+R_3}$ the departures are out of systems 4 and 3 only.
- d. With probability $\frac{[R_2-R_4]^+}{R_2+R_3}$ there is departure out of system 2 only.
- e. With probability $\frac{R_3-R_1-[R_4+R_2]^+}{R_2+R_3}$ only a departure out of system 3 occurs.
- f. With probability $\frac{R_2}{R_2+R_3}$ no transitions occur.

Similar transitions apply if $R_4 = \min\{R_i\}$ by switching the roles of systems 1 and 4.

Two observations can be made regarding these transitions: a) Every departure from systems 1 or 4 is coupled with a departure from systems 2 or 3, and b) Whenever $Y_1 = Y_i$ for some i , a departure from system 1 is coupled with a departure in system i . Therefore, both 1. and 2. are satisfied at time $n + 1$.

- (ii) Suppose $\eta \geq \mu$. We show that if $S_1 \leq S_2 \leq S_3 \leq S_4$ with $S_1 + S_4 = S_2 + S_3$, then there exist versions of the head-count processes: Y_1, \dots, Y_4 , respectively, such that for all $t \geq 0$

1. $Y_1(t) + Y_4(t) \leq Y_2(t) + Y_3(t)$. a.s., and
2. $Y_1(t) \leq \min\{Y_2(t), Y_3(t), Y_4(t)\}$, a.s.

Conversely to the previous case, suppose that (A8) holds, then one can show that

$$R_1(n) + R_4(n) \geq R_2(n) + R_3(n), \quad (\text{A10})$$

In addition, it can be easily shown that if $Y_1 = Y_i$ for some $i \neq 1$, then $R_1 \geq R_i$. Provided that the above is true, we now assume that (A8) and 2. hold at time n and show that 1. and 2. will also hold at time $n + 1$. Again, we define reneging and service completion together as departures. If $R_2 = \min\{R_i\}$ (n is omitted for brevity), then departures will be defined as follows:

- a. With probability $\frac{R_2}{R_1+R_4}$ we have departures out of all systems.
- b. With probability $\frac{\min\{R_3, R_1\} - R_2}{R_1+R_4}$ we have departures out of systems 1 and 3 only.
- c. With probability $\frac{[R_1 - R_3]^+}{R_1+R_4}$ we have a departure out of system 1 only.
- d. With probability $\frac{[R_3 - R_1]^+}{R_1+R_4}$ we have departures out of systems 3 and 4 only.
- e. With probability $\frac{R_4 - R_2 - [R_3 - R_1]^+}{R_1+R_4}$ we have a departure out of system 4 only.
- f. With probability $\frac{R_2}{R_1+R_4}$ no transitions occur.

If $R_3 = \min\{R_i\}$ similar transitions occur, switching the roles of systems 2 and 3.

If, on the other hand, $R_1 = \min\{R_i\}$, then from (A10), $R_4 = \max\{R_i\}$. In this case the coupling works as follows:

- a. With probability $\frac{R_1}{R_1+R_4}$ we have departures out of all systems.
- b. With probability $\frac{R_2 - R_1}{R_1+R_4}$ we have departures out of systems 2 and 4 only.
- c. With probability $\frac{R_3 - R_1}{R_1+R_4}$ we have departures out of systems 3 and 4 only.
- d. With probability $\frac{R_4 + R_1 - (R_2 + R_3)}{R_1+R_4}$ we have a departure out of system 4 only.
- e. With probability $\frac{R_1}{R_1+R_4}$ no transitions occur.

If $R_4 = \min\{R_i\}$ then we have similar transitions by switching the roles of systems 1 and 4.

It is easy to verify that 1. and 2. will hold at time $n + 1$ using these transitions. \square

References

- G. Koole and A. Pot. 2006. A note on profit maximization and monotonicity for inbound call centers. Working paper.
- Shaked, M. and J.G. Shanthikumar. 1988. Stochastic Convexity and Its Applications, *Adv. Appl. Probability* 20 427-446.