# Online Appendices: Dimensioning On-Demand Vehicle Sharing Systems

Saif Benjaafar[†], Shining Wu[‡], Hanlin Liu[§], Einar Bjarki Gunnarsson[†]

[†]Department of Industrial & Systems Engineering, University of Minnesota, Twin Cities, USA

[‡]Department of Logistics & Maritime Studies, Hong Kong Polytechnic University, Hong Kong

[§]Division of Information Systems and Management Engineering, College of Business, Southern University of Science and Technology, Shenzhen, China

saif@umn.edu, sn.wu@polyu.edu.hk, liuhl@sustech.edu.cn, gunna042@umn.edu

# Appendix A Proofs

In the section, we provide proofs for our main results.

To prove Lemma 2, it is sufficient to prove the following more general lemma.

**Lemma 5.** *(i)* $\Delta\alpha(K) \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$ *holds for all $K$, and the first inequality is strict for $K > 1$.*

*(ii)* $\Delta\alpha(K) \leq \frac{\mu}{(K+N-1)\mu+\Lambda[1-\alpha(K-1)]}$ *holds for all $K$ and is strict for $K > 1$.*

*(iii)* $0 < \Delta\alpha(K) < \Delta\alpha(K-1)$. *That is, $\alpha(K)$ is increasing concave in $K$.*

*(iv)* $\Delta\alpha(K) > \frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{\left\{(K+N)+\frac{\Lambda}{\mu}[1-\alpha(K)]\right\}^2}$.

**Proof of Lemma 5.** First, note that from (8), we can obtain

$$\Delta\alpha(K) = \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)] + K\frac{\Lambda}{\mu}[\Delta\alpha(K-1)]}{[(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))][(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))]},$$

which also corresponds to equation (12) in the main text. We will use this result in the course of this proof.

We prove the results (i)–(iv) in sequence.

(i) We prove $\Delta\alpha(K) \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$ by induction. Note that the inequalities are shown to be strict for $K > 1$ in our proof.

- When $K = 1$, $\Delta\alpha(1) = \alpha(1) - \alpha(0) = \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$.
- Suppose that $\Delta\alpha(k) \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$ holds for $k \leq K-1$, we then prove that the inequality strictly holds for $k = K$.

  By plugging $\alpha(K-1) - \alpha(K-2) < \frac{\mu}{\Lambda}$ into (12), we have

  $$\alpha(K) - \alpha(K-1) < \frac{1}{(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))} \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda},$$

  where the second inequality holds because $\alpha(k) \leq \frac{k\mu}{\Lambda}$ for any $k \geq 0$.

(ii) Second, note that the inequality $\Delta\alpha(K) < \frac{\mu}{\Lambda}$ guarantees that $(K+N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)]$ increases in $K$. Therefore,

$$\begin{aligned}
\alpha(K) - \alpha(K-1) &= \frac{K}{(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))} - \frac{K-1}{(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))} \\
&< \frac{K}{(K+N-1)\mu+\frac{\Lambda}{\mu}(1-\alpha(K-1))} - \frac{K-1}{(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))} \\
&= \frac{\mu}{(K+N-1)\mu+\Lambda(1-\alpha(K-1))},
\end{aligned}$$

for $K \geq 2$ where both $\alpha(K)$ and $\alpha(K-1)$ can be expressed by the recursive equation (8). Furthermore, it is obvious that the above inequality also holds for $K = 1$ since $\alpha(0) = 0$.

(iii) Next, we prove result (iii) by induction. On the one hand, by (12) and the fact that $\alpha(1) > \alpha(0)$, a simple induction proves $\alpha(K) > \alpha(K-1)$, $\forall K$. That is, $\Delta\alpha(K) > 0$. On the other hand, it is easy to verify $\Delta\alpha(2) < \Delta\alpha(1)$. Assume that $\Delta\alpha(k) < \Delta\alpha(k-1)$ holds for $k \leq K-1$, we then prove $\Delta\alpha(K) < \Delta\alpha(K-1)$. By (12), we know that $\Delta\alpha(K) < \Delta\alpha(K-1)$ holds if and only if

$$(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)] < \Delta\alpha(K-1)\left\{\left[(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]-K\frac{\Lambda}{\mu}\right\}.$$

Note that

$$\begin{aligned}&\left[(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]-K\frac{\Lambda}{\mu}\\ =&\left[(K+N-3)+\frac{\Lambda}{\mu}(1-\alpha(K-3))+2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]-K\frac{\Lambda}{\mu}.\end{aligned}$$

Thus, the condition is satisfied if

$$\begin{aligned}&\Delta\alpha(K-1)\\ >&\frac{(N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))+\Delta\alpha(K-1)\left\{K\frac{\Lambda}{\mu}-\left[2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]\right\}}{\left[(K+N-3)+\frac{\Lambda}{\mu}(1-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]}\\ =&\frac{(N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-2))+\Delta\alpha(K-1)\left\{(K-1)\frac{\Lambda}{\mu}-\left[2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]\right\}}{\left[(K+N-3)+\frac{\Lambda}{\mu}(1-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]},\end{aligned}$$

which holds according to (12) because $\Delta\alpha(K-2) > \Delta\alpha(K-1)$ by the induction assumption and that $[2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))] > 0$ by the result (i) in this proposition.

(iv) Lastly, according to the result (i), $(K+N)\mu+\Lambda[1-\alpha(K)]$ increases in $K$ because $\Delta\alpha(K) \leq \frac{\mu}{\Lambda}$. Therefore, both $\alpha(K)$ and $(K+N)\mu+\Lambda[1-\alpha(K)]$ increase in $K$, and hence equation (12) yields

$$\begin{aligned}\alpha(K)-\alpha(K-1) &> \frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{[(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))][(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))]}\\ &> \frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{[(K+N)+\frac{\Lambda}{\mu}(1-\alpha(K))][(K+N)+\frac{\Lambda}{\mu}(1-\alpha(K))]}.\end{aligned}$$

□

**Proof of Proposition 1.** From (8), we have

$$K = \frac{\Lambda}{\mu}\alpha(K)+(N-1)\frac{\alpha(K)}{1-\alpha(K)}+\frac{\Lambda}{\mu}\frac{\alpha(K)}{1-\alpha(K)}\Delta\alpha(K)$$

Because $K(\alpha)$ is the smallest number that satisfies $\alpha(K) \geq \alpha$, we know $\alpha\big(K(\alpha)-1\big) < \alpha \leq \alpha\big(K(\alpha)\big)$. Therefore,

$$\begin{aligned}
K(\alpha) &= \frac{\Lambda}{\mu}\alpha\big(K(\alpha)\big) + (N-1)\frac{\alpha\big(K(\alpha)\big)}{1-\alpha\big(K(\alpha)\big)} + \frac{\Lambda}{\mu}\frac{\alpha\big(K(\alpha)\big)}{1-\alpha\big(K(\alpha)\big)}\Delta\alpha\big(K(\alpha)\big) \\
&\geq \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\Delta\alpha\big(K(\alpha)\big), \qquad (20) \\
K(\alpha)-1 &= \frac{\Lambda}{\mu}\alpha\big(K(\alpha)-1\big) + (N-1)\frac{\alpha\big(K(\alpha)-1\big)}{1-\alpha\big(K(\alpha)-1\big)} + \frac{\Lambda}{\mu}\frac{\alpha\big(K(\alpha)-1\big)}{1-\alpha\big(K(\alpha)-1\big)}\Delta\alpha\big(K(\alpha)-1\big) \\
&< \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\Delta\alpha\big(K(\alpha)-1\big). \qquad (21)
\end{aligned}$$

First, it is easy to see

$$K(\alpha) \geq \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha},$$

since $\Delta\alpha(K) > 0$. The inequality is strict if $K(\alpha) \geq 1$, i.e., $\alpha > 0$.

Second, by substituting $\Delta\alpha(K) < \frac{\mu}{\Lambda}$ into the right hand side of (21), we obtain

$$K(\alpha) < \frac{\Lambda}{\mu}\alpha + N\frac{\alpha}{1-\alpha} + 1.$$

$\square$

Note that equation (12) also describes a recursive relationship between $\Delta\alpha(K)$ and $\Delta\alpha(K-1)$. A series of bounds on $\Delta\alpha(K)$ can be obtained if we expand (12) for multiple times and bound $\Delta\alpha(K-s)$ according to Lemma 5. We introduce the following result as a corollary of Lemma 5.

**Corollary 2.** *Let $\tilde{\eta}_0(K) := 0$ and $\tilde{\zeta}_0(K) := \frac{\mu}{(K+N-1)\mu+\Lambda(1-\alpha(K-1))}$ for all $K \geq 1$. Define $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ iteratively for $s = 1, 2, \cdots, K-1$, where $K \geq 2$ as follows.*

$$\begin{aligned}
\tilde{\eta}_s(K) &= \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)] + K\frac{\Lambda}{\mu}\tilde{\eta}_{s-1}(K-1)}{\{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]\}\{(K+N-2)+\frac{\Lambda}{\mu}[1-\alpha(K-2)]\}}, \\
\tilde{\zeta}_s(K) &= \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)] + K\frac{\Lambda}{\mu}\tilde{\zeta}_{s-1}(K-1)}{\{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]\}\{(K+N-2)+\frac{\Lambda}{\mu}[1-\alpha(K-2)]\}}.
\end{aligned}$$

*Then, $\tilde{\eta}_s(K) < \Delta\alpha(K) \leq \tilde{\zeta}_s(K)$ for all $0 \leq s \leq K-1$. Furthermore, $\tilde{\eta}_s(K) > \tilde{\eta}_{s-1}(K)$ and $\tilde{\zeta}_s(K) < \tilde{\zeta}_{s-1}(K)$.*

**Proof of Corollary 2.** By iterating equation (12) multiple times on its right hand side, we are able to express $\Delta\alpha(K)$ by $\Delta\alpha(K-s)$, where the last iteration has terms $(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-s)] + (K-s+1)\frac{\Lambda}{\mu}\Delta\alpha(K-s)$ in its numerator. Noting that $(K-s)\Delta\alpha(K-s) \leq \alpha(K-s)$ by result (iii) of Lemma 5, we know $\frac{\Lambda}{\mu}[1-\alpha(K-s)] + (K-s+1)\frac{\Lambda}{\mu}\Delta\alpha(K-s) \leq \frac{\Lambda}{\mu} + \frac{\Lambda}{\mu}\Delta\alpha(K-s) \leq \frac{\Lambda}{\mu} + \tilde{\zeta}_0(K-s)$. By replacing the term $(K-s+1)\frac{\Lambda}{\mu}\Delta\alpha(K-s)$ with 0 on the right hand side of the iteration, we have $\Delta\alpha(K) > \tilde{\eta}_s(K)$. By replacing the term $\frac{\Delta}{\mu}[1-\alpha(K-s)]+(K-s+1)\frac{\Delta}{\mu}\Delta\alpha(K-s)$ with $\frac{\Delta}{\mu}+\tilde{\zeta}_0(K-s)$ on the right hand side of the iteration, we have $\Delta\alpha(K) \le \tilde{\zeta}_s(K)$.

According to Lemma 5, $\tilde{\eta}_1(K) > 0 = \tilde{\eta}_0(K)$ and

$$\begin{aligned}\tilde{\zeta}_1(K) &= \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha(K-1)]}{\{(K+N-1)+\frac{\Delta}{\mu}[1-\alpha(K-1)]\}\{(K+N-2)+\frac{\Delta}{\mu}[1-\alpha(K-2)]\}} \\ &< \frac{1}{\{(K+N-1)+\frac{\Delta}{\mu}[1-\alpha(K-1)]\}} = \tilde{\zeta}_0(K)\end{aligned}$$

for $K \ge 2$. By the definition of $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$, we can easily prove $\tilde{\eta}_s(K) > \tilde{\eta}_{s-1}(K)$ and $\tilde{\zeta}_s(K) < \tilde{\zeta}_{s-1}(K)$ by induction. □

Although $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ bound $\Delta\alpha(K)$ in general, they are not useful bounds for $\Delta\alpha\big(K(\alpha)\big)$ because the value of $K(\alpha)$ is unknown. Thus, we replace the value $K$ in the definitions of $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ with its upper bound $U$ and lower bound $L$ to yield $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$, whose expressions do not involve $K$ and hence are now practical bounds for $\Delta\alpha(K)$. Recall that the definitions of $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$ are presented in the main text. Here we prove the following proposition.

**Lemma 6.** *If $L \le K(\alpha) \le U$, then $\tilde{\eta}_s\big(K(\alpha)-t\big) \ge \eta_s^t(L,U,\alpha)$ and $\tilde{\zeta}_s\big(K(\alpha)-t\big) \le \zeta_s^t(L,U,\alpha)$, for all $t \ge 0$ and $0 \le s \le L-1$.*

**Proof of Lemma 6.** Note that $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$ are defined in similar ways as $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ in Corollary 2 with minor differences in the numerators and denominators. While the denominator of $\tilde{\eta}_s(K)$ is updated according to $K$, that of $\eta_s^t(L,U,\alpha)$ depends only on $U$ and $\alpha$.

By the fact that $L \le K(\alpha) \le U$ and $\alpha\big(K(\alpha)-1\big) < \alpha \le \alpha\big(K(\alpha)\big)$, we prove the results by induction for $s = 0, 1, 2, \cdots, L-1$.

- First, $s = 0$. We know that $\tilde{\eta}_0\big(K(\alpha)-t\big) = \eta_0^t(L,U,\alpha) = 0$ and $\tilde{\zeta}_0\big(K(\alpha)-t\big) \le \zeta_0^t(L,U,\alpha)$, $\forall t \ge 0$ by result (i) of Lemma 5.

- Suppose that we have proved the result for indices up to $s-1$. Then, by the definition of $\tilde{\eta}_s$, we have

$$\begin{aligned}
\tilde{\eta}_s(K(\alpha)-t) =& \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\eta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N-t-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]\Big\}\Big\{(K(\alpha)+N-t-2)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-2\big)]\Big\}}\\
>& \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\eta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)\big)]\Big\}^2}\\
>& \frac{(N-1)+\frac{\Delta}{\mu}(1-\alpha)+(L-t)\frac{\Delta}{\mu}\tilde{\eta}_{s-1}(K(\alpha)-t-1)}{\Big[(U+N)+\frac{\Delta}{\mu}(1-\alpha)\Big]^2}\\
\geq& \frac{(N-1)+\frac{\Delta}{\mu}(1-\alpha)+(L-t)\frac{\Delta}{\mu}\eta_{s-1}^{t+1}(L,U,\alpha)}{\Big[(U+N)+\frac{\Delta}{\mu}(1-\alpha)\Big]^2} = \eta_s^t(L,U,\alpha),
\end{aligned}$$

where the first inequality holds because $K+\frac{\Delta}{\mu}[1-\alpha(K)]$ increases in $K$ (by result (i) of Lemma 5), the second inequality follows from the fact that $L \leq K(\alpha) \leq U$, $\alpha\big(K(\alpha)-1\big) < \alpha \leq \alpha\big(K(\alpha)\big)$, and the last inequality holds by the induction hypothesis.

On the other hand, by the definition of $\tilde{\zeta}_s$,

$$\begin{aligned}
\tilde{\zeta}_s(K(\alpha)-t) =& \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\zeta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N-t-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]\Big\}\Big\{(K(\alpha)+N-t-2)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-2\big)]\Big\}}\\
<& \frac{(N-1)+\frac{\Delta}{\mu}\Big\{1-\alpha\big(K(\alpha)\big)+\frac{t+1}{K(\alpha)+N-t-1+\frac{\Delta}{\mu}\big[1-\alpha(K(\alpha)-t-1)\big]}\Big\}+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\zeta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N-t-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]\Big\}\Big\{(K(\alpha)+N-t-2)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-2\big)]\Big\}}\\
<& \frac{(N-1)+\frac{\Delta}{\mu}\big\{1-\alpha+\frac{t+1}{L+N-t-1+\frac{\Delta}{\mu}(1-\alpha)}\big\}+\big(U-t\big)\frac{\Delta}{\mu}\tilde{\zeta}_{s-1}(K(\alpha)-t-1)}{\Big\{(L+N-t-1)+\frac{\Delta}{\mu}(1-\alpha)\Big\}\Big\{(L+N-t-2)+\frac{\Delta}{\mu}(1-\alpha)\Big\}}\\
\leq& \frac{(N-1)+\frac{\Delta}{\mu}\big\{1-\alpha+\frac{t+1}{L+N-t-1+\frac{\Delta}{\mu}(1-\alpha)}\big\}+\big(U-t\big)\frac{\Delta}{\mu}\zeta_{s-1}^{t+1}(L,U,\alpha)}{\Big\{(L+N-t-1)+\frac{\Delta}{\mu}(1-\alpha)\Big\}\Big\{(L+N-t-2)+\frac{\Delta}{\mu}(1-\alpha)\Big\}}\\
=& \zeta_s^t(L,U,\alpha),
\end{aligned}$$

where the second inequality follows from the fact that $L \leq K(\alpha) \leq U$, $\alpha\big(K(\alpha)-1\big) < \alpha \leq \alpha\big(K(\alpha)\big)$ and the last inequality holds by the induction hypothesis. The first inequality holds because

$$\begin{aligned}
1-\alpha\big(K-t-1\big) &= 1-\alpha(K)+\alpha(K)-\alpha(K-t-1) \leq 1-\alpha(K)+(t+1)\Delta\alpha(K-t)\\
&\leq 1-\alpha(K)+(t+1)\frac{1}{K-t+N-1+\frac{\Delta}{\mu}[1-\alpha(K-t-1)]}
\end{aligned}$$

from results (ii) and (iii) of Lemma 5.

$\square$

**Proof of Lemma 3.** The results follow directly from Corollary 2 and Lemma 6. $\square$

**Proof of Proposition 2.** The results follow directly from Lemma 3 and inequalities (20) and (21). $\square$

**Proof of Corollary 1.** By applying Proposition 2 to the bounds $(L_0, U_0)$ established in Proposition 1, we have $L_s < K(\alpha) < U_s$.

To prove $L_s \geq L_{s-1}$ and $U_s \leq U_{s-1}$, it is sufficient to show that $\eta_s^t(L_0, U_0, \alpha) > \eta_{s-1}^t(L_0, U_0, \alpha)$ and $\zeta_s^t(L_0, U_0, \alpha) > \zeta_{s-1}^t(L_0, U_0, \alpha)$ for $s = 0, 1, \cdots$.

For the monotonicity result of $\eta_s^t$, we can prove for any pair of positive bounds as argument of $\eta_s^t(L, U, \alpha)$. It is obvious that $\eta_1^t(L, U, \alpha) > 0 = \eta_0^t(L, U, \alpha)$, $\forall t$. According to the definition of $\eta_s^t(L, U, \alpha)$, we know $\eta_s^t(L, U, \alpha) > \eta_{s-1}^t(L, U, \alpha)$ for all $0 \leq s \leq L - 1$, which is an analogue to the result $\tilde{\eta}_s(K) > \tilde{\eta}_{s-1}(K)$ in Corollary 2.

For the monotonicity result of $\zeta_s^t$, we only prove the case with $(L_0, U_0)$ as argument and for certain $s$'s. Again, according to the definition of $\zeta_s^t(L, U, \alpha)$, it is sufficient to prove $\zeta_1^t(L_0, U_0, \alpha) < \zeta_0^t(L_0, U_0, \alpha)$. We then examine the condition that can guarantee this inequality. Note that

$$\zeta_1^t(L, U, \alpha) = \frac{(N-1) + \frac{\Lambda}{\mu}[1 - \alpha + \frac{t+1}{(L_0+N-t-1)+\frac{\Lambda}{\mu}(1-\alpha)}] + (U_0 - t)\frac{\Lambda}{\mu}\zeta_0^{t+1}(L_0, U_0, \alpha)}{[(L_0 + N - t - 1) + \frac{\Lambda}{\mu}(1-\alpha)][(L_0 + N - t - 2) + \frac{\Lambda}{\mu}(1-\alpha)]} < \frac{\mu}{\Lambda} = \zeta_0^t(K, L, \alpha)$$

$$\Leftrightarrow \frac{\Lambda}{\mu}(U_0 - L_0) + \frac{(\frac{\Lambda}{\mu})^2(t+1)}{(L_0 + N - t - 1) + \frac{\Lambda}{\mu}(1-\alpha)} < [(L_0 + N - t - 1) + \frac{\Lambda}{\mu}(1-\alpha)][(L_0 + N - t - 2) + \frac{\Lambda}{\mu}(1-\alpha) - \frac{\Lambda}{\mu}].$$

By substituting the values of $L_0$ and $U_0$ into the above inequality, we know that the condition

$$\frac{\Lambda}{\mu}\left[\frac{1}{1-\alpha} + 1 + \frac{\frac{\Lambda}{\mu}(t+1)}{\frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha} - t}\right] < \left(\frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha} - t\right)\left(\frac{N-1}{1-\alpha} - t - 1\right)$$

is needed to guarantee $\zeta_1^t(L_0, U_0, \alpha) < \zeta_0^t(L_0, U_0, \alpha)$. If $t + 1 \leq \frac{1}{6}\frac{N-1}{1-\alpha}$, simple algebra shows that the left hand side is no greater than $\frac{\Lambda}{\mu}\frac{2+\frac{N-1}{6}}{1-\alpha} - \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}$ and the right hand side is greater than $\frac{\Lambda}{\mu}\frac{5}{6}\frac{N-1}{1-\alpha} + (\frac{5}{6}\frac{N-1}{1-\alpha})^2$. When $N \geq 4$, the inequality holds regardless of the value of $\frac{\Lambda}{\mu}$ and $\alpha$ because $2 + \frac{N-1}{6} \leq \frac{5}{6}(N-1)$. Noting that $s + t$ is constant in the iteration, the condition $t + 1 \leq \frac{1}{6}\frac{N-1}{1-\alpha}$ requires $s + 1 \leq \frac{1}{6}\frac{N-1}{1-\alpha}$ as well. Therefore, $\zeta_s^t(L_0, U_0, \alpha) < \zeta_{s-1}^t(L_0, U_0, \alpha)$ and hence $U_s < U_{s-1}$ if $s < \frac{1}{6}\frac{N-1}{1-\alpha}$ and $N \geq 4$. $\square$

As one can see, $\eta_s^t$ and $\zeta_s^t$ are defined using recursive equations that are similar to (12) for $\Delta\alpha(K-t)$'s, $t = 0, 1, \cdots, K$ except that the unknown variables in (12) are replaced by their known bounds, e.g, $K(\alpha)$ by $L$ or $U$, $\alpha(K)$ by $\alpha$, $\alpha(K-t)$ by $\alpha$ or $\alpha - \frac{t}{L+N-t+\frac{\Lambda}{\mu}(1-\alpha)}$, etc. On the one hand, the expressions of $\eta_s^t$ and $\zeta_s^t$ better resemble the recursive formulation of the actual $\Delta\alpha(K)$, leading to better approximations, as more iterations are applied. On the other hand, the replacement of the unknown actual variables with their bounds introduces additional errors with each iteration, which may lead to a deterioration in the quality of the generated bounds. In particular, such relaxation errors could be significant for the upper bounds $\zeta_s^t$'s because one of the approximate term, $\frac{(t+1)\Lambda/\mu}{[L+N-t-1+\frac{\Lambda}{\mu}(1-\alpha)]^2[L+N-t-2+\frac{\Lambda}{\mu}(1-\alpha)]}$, is increasing convex in $t$ (i.e., larger and larger relaxation errors are introduced with each iteration). We observe that the net effect is that the upper bounds $\zeta_s^1(L,U,\alpha)$'s may deteriorate with additional iterations when many iterations have already been applied and the bound is already very close to the actual $\Delta\alpha\big(K(\alpha)\big)$. This does not happen to the lower bounds because the unknown variables in their expressions are replaced by the parameters that remain constant over iterations. Thus, the $\eta_s^0(L,U,\alpha)$'s always increase (improve) over iterations. An illustration of this effect is presented in Figure 3.

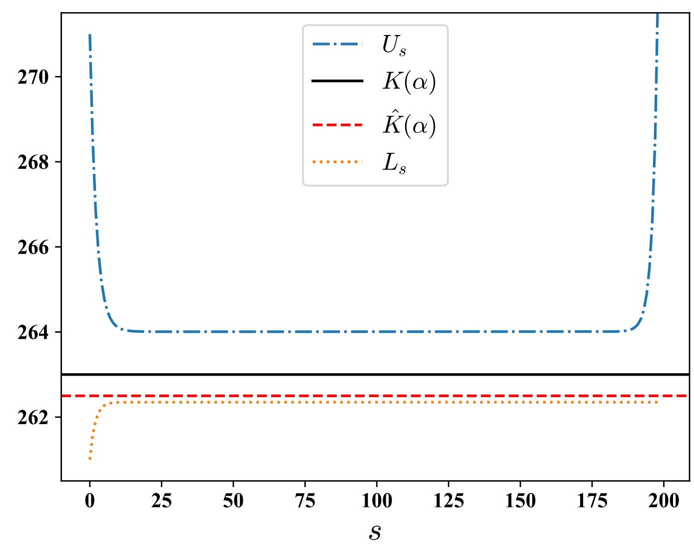



Figure 3: The lower and upper bounds as a function of the number of iterations ($\alpha = 0.9$, $\Lambda = 200$, $\mu = 1$, and $N = 10$)

**Proof of Proposition 3.** According to the definition of $L_s$ and $U_s$, it is sufficient to prove the following results.

1. For fixed $N$ and any finite $s,t \geq 1$, $\lim\limits_{\Lambda\to 0^+} \frac{\Lambda}{\mu}\eta_s^0(L_0,U_0,\alpha) = \lim\limits_{\Lambda\to 0^+} \frac{\Lambda}{\mu}\zeta_s^1(L_0,U_0,\alpha) = 0$.

2. For fixed $N$ and any finite $s,t \geq 0$, $\lim\limits_{\Lambda\to\infty} \frac{\Lambda}{\mu}\eta_s^t(L_0,U_0,\alpha) = 1-\alpha^s$, and $\lim\limits_{\Lambda\to\infty} \frac{\Lambda}{\mu}\zeta_s^t(L_0,U_0,\alpha) = 1$.

3. For fixed $\Lambda$ and any finite $s,t \geq 1$, $\lim\limits_{N\to\infty} \frac{\Lambda}{\mu}\eta_s^t(L_0,U_0,\alpha) = \lim\limits_{\Lambda\to 0^+} \frac{\Lambda}{\mu}\zeta_s^t(L_0,U_0,\alpha) = 0$.

4. For fixed $\lambda = \frac{\Lambda}{N}$ and any finite $s, t \geq 0$,

$$\lim_{N\to\infty} \frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha) = \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)}\left[1 - \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}}\right)^s\right], \text{ and}$$

$$\lim_{N\to\infty} \frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha) = \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)} + \frac{\frac{1}{1-\alpha}}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)}\left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}}\right)^s.$$

5. For fixed $\Lambda > 0$, $N > 1$, and any finite $s, t \geq 2$, $\lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = \lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = 0$. For $s = 1$, $\lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\eta_1^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = 0$ and $\lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\zeta_1^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = \frac{\Lambda}{\mu}\frac{N}{(N-1)^2}$.

These results can be proven by induction for $s = 1, 2, \cdots$. Noting that

$$L_0 - t = \frac{\Lambda}{\mu}\alpha + \frac{(N-1)\alpha}{1-\alpha} - t, \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha} - t,$$

$$U_0 - t = \frac{\Lambda}{\mu}\alpha + \frac{N\alpha}{1-\alpha} + 1 - t, \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + \frac{N}{1-\alpha} + 1,$$

we know that (1) as $\Lambda \to 0$, $\lim\limits_{\Lambda\to 0} L_0 = 0$, $\lim\limits_{\Lambda\to 0} U_0 = 1$; (2) as $\Lambda \to \infty$,

$$L_0 - t = \frac{\Lambda}{\mu}\alpha + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + O(1),$$

$$U_0 - t = \frac{\Lambda}{\mu}\alpha + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + O(1);$$

(3) as $N \to \infty$,

$$L_0 - t = \frac{N\alpha}{1-\alpha} + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N}{1-\alpha} + O(1),$$

$$U_0 - t = \frac{N\alpha}{1-\alpha} + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N}{1-\alpha} + O(1);$$

(4) as $N \to \infty$ and $\Lambda = N\lambda$,

$$L_0 - t = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right)\alpha + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right) + O(1),$$

$$U_0 - t = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right)\alpha + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right) + O(1);$$

and (5) as $\alpha \to 1$,

$$L_0 - t = \frac{(N-1)\alpha}{1-\alpha} + O(1) = \frac{N-1}{1-\alpha} + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N-1}{1-\alpha} + O(1),$$

$$U_0 - t = \frac{N\alpha}{1-\alpha} + O(1) = \frac{N}{1-\alpha} + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N}{1-\alpha} + O(1).$$

The proofs for regimes 1, 2, 3, and 5 can be completed using induction by applying the above limits to the definition of $\eta_s^t$ and $\zeta_s^t$. Here, we only provide the detailed steps for regime 4, the case that requires most involved expressions.

- For $s = 0$, $\frac{\Lambda}{\mu}\eta_0^t(L_0, U_0, \alpha) = 0$ and $\frac{\Lambda}{\mu}\zeta_0^t(L_0, U_0, \alpha) = 1$, which satisfy the induction hypothesis for all $t$.
- Suppose that the results hold for subscript $0, 1, \cdots, s-1$ for all $t$, then

$$\begin{aligned}
\lim_{N\to\infty}\frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha) &= \frac{\frac{\lambda}{\mu}\left\{1 + \frac{\lambda}{\mu}(1-\alpha) + (\frac{\lambda}{\mu} + \frac{1}{1-\alpha})\alpha\frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha}+\frac{\lambda}{\mu}(1-\alpha)}\left[1 - \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha}+\frac{\lambda}{\mu}}\right)^{s-1}\right]\right\}}{(\frac{\lambda}{\mu} + \frac{1}{1-\alpha})^2} \\
&= \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)}\left[1 - \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}}\right)^{s}\right], \\
\lim_{N\to\infty}\frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha) &= \frac{\frac{\lambda}{\mu}\left\{1 + \frac{\lambda}{\mu}(1-\alpha) + (\frac{\lambda}{\mu} + \frac{1}{1-\alpha})\alpha\left[\frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}(1-\alpha)} + \frac{\frac{1}{1-\alpha}}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}(1-\alpha)}\left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}}\right)^{s-1}\right]\right\}}{(\frac{\lambda}{\mu} + \frac{1}{1-\alpha})^2} \\
&= \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)} + \frac{\frac{1}{1-\alpha}}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)}\left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}}\right)^{s}.
\end{aligned}$$

□

**Proof of Proposition 4.** We prove the result for all $s = 0, \ldots, L_0 - 1$ where $(L_s, U_s)$ is well-defined in Corollary 1. We only consider the non-trivial case where $L_0 \geq 1$, i.e., $(\frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha})\alpha \geq 1$. By comparing the expressions of (13), (14), and (15), it suffices to prove that

$$\eta_s^t(L_0, U_0, \alpha) < \frac{1}{\frac{N}{(1-\alpha)^2} + \frac{\Lambda}{\mu}} < \zeta_s^{t+1}(L_0, U_0, \alpha) \tag{22}$$

holds for all $s \geq 0$ and $t \geq 0$. We prove by induction on progressing index $s$ as follows.

- For $s = 0$, $\eta_0^t(L_0, U_0, \alpha) = 0 < \left[\frac{N}{(1-\alpha)^2} + \frac{\Lambda}{\mu}\right]^{-1} < \frac{\mu}{\Lambda} = \eta_0^{t+1}(L_0, U_0, \alpha)$, which satisfies the induction hypothesis.
- Suppose that (22) holds for subscripts $0, 1, \cdots, s-1$ for all $t$, we then examine whether it holds for subscript $s$. We know by definition that

$$\begin{aligned}
\eta_s^0(L_0, U_0, \alpha) &= \frac{(N-1) + \frac{\Lambda}{\mu}(1-\alpha) + L_0\frac{\Lambda}{\mu}\eta_{s-1}^1(L_0, U_0, \alpha)}{(\frac{N}{1-\alpha} + 1 + \frac{\Lambda}{\mu})^2} \quad \text{and} \\
\zeta_s^1(L_0, U_0, \alpha) &= \frac{(N-1) + \frac{\Lambda}{\mu}[1 - \alpha + \frac{2}{(L_0+N-2)+\frac{\Lambda}{\mu}(1-\alpha)}] + (U_0 - 1)\frac{\Lambda}{\mu}\zeta_{s-1}^2(L_0, U_0, \alpha)}{(\frac{N-1}{1-\alpha} - 1 + \frac{\Lambda}{\mu}) \cdot (\frac{N-1}{1-\alpha} - 2 + \frac{\Lambda}{\mu})}.
\end{aligned}$$

By simple algebra, the former is smaller than $\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}$ if

$$\left[L_0\frac{\Lambda}{\mu}\eta^1_{s-1}(L_0,U_0,\alpha)-1\right]\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]<\left(\frac{\Lambda}{\mu}+\frac{N}{1-\alpha}\right)\left(\frac{\Lambda}{\mu}\alpha+2\right)+1;$$

the latter is greater than $\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}$ if

$$(U_0-1)\frac{\Lambda}{\mu}\zeta^2_{s-1}(L_0,U_0,\alpha)\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]+\frac{2\frac{\Lambda}{\mu}\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]}{\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}-1}>\left(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\right)\frac{\Lambda}{\mu}\alpha-\left(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\right)(3+\frac{1}{\alpha})+2.$$

Both of these conditions hold because

- $L_0\frac{\Lambda}{\mu}\eta^1_{s-1}(L_0,U_0,\alpha)\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]<L_0\frac{\Lambda}{\mu}=\left(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\right)\frac{\Lambda}{\mu}\alpha$ by the induction hypothesis for subscript $s-1$.
- $(U_0-1)\frac{\Lambda}{\mu}\zeta^2_{s-1}(L_0,U_0,\alpha)\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]>(U_0-1)\frac{\Lambda}{\mu}=\left(\frac{\Lambda}{\mu}+\frac{N}{1-\alpha}\right)\frac{\Lambda}{\mu}\alpha$ by the induction hypothesis for subscript $s-1$.
- $-\left(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\right)(3+\frac{1}{\alpha})+2<-\frac{1}{\alpha}(3+\frac{1}{\alpha})+2<0$ because $(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha})\alpha\geq 1$ in the non-trivial case.

Therefore, we obtain $\eta^0_s(L_0,U_0,\alpha)<\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}<\zeta^1_s(L_0,U_0,\alpha)$. Since $\eta^t_s(L_0,U_0,\alpha)$ decreases in $t$ and $\zeta^t_s(L_0,U_0,\alpha)$ increases in $t$, we further obtain $\eta^t_s(L_0,U_0,\alpha)<\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}<\zeta^{t+1}_s(L_0,U_0,\alpha)$ for all $t\geq 0$. The induction is completed.

□

**Proof of Proposition 5.** The proof is trivial according to (15). □

**Proof of Proposition 6.** By the monotonicity of $\alpha(K)$ and (8), we know that $\lim_{K\to\infty}\alpha(K)$ exists and equals to 1. By (8), we know that

$$1-\alpha(K)=\frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}. \tag{23}$$

<u>If $N>1$</u>, then

$$\lim_{K\to\infty}\left\{[1-\alpha(K)]\frac{K}{N-1}\right\}=\lim_{K\to\infty}\left\{\frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{N-1}\frac{K}{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}\right\}=1.$$

If $\underline{N=1}$, by (23), we know

$$\frac{1}{1-\alpha(K)} = 1 + \frac{K(\frac{\Lambda}{\mu})^{-1}}{1-\alpha(K-1)}, \text{ for all } K, \text{ and} \tag{24}$$

$$\frac{1}{1-\alpha(K)} + \frac{\frac{\Lambda}{\mu}}{K-\frac{\Lambda}{\mu}} = K(\frac{\Lambda}{\mu})^{-1}\left(\frac{1}{1-\alpha(K-1)} + \frac{\frac{\Lambda}{\mu}}{K-\frac{\Lambda}{\mu}}\right), \text{ for all } K \neq \frac{\Lambda}{\mu}. \tag{25}$$

On the one hand, let $\underline{\omega}(K) := K(\frac{\Lambda}{\mu})^{-1}\underline{\omega}(K-1)$ be recursively defined with $\underline{\omega}(1) := \frac{1}{1-\alpha(1)} = (\frac{\Lambda}{\mu})^{-1}(1+\frac{\Lambda}{\mu})$. Then, $\underline{\omega}(K) = (K!)(\frac{\Lambda}{\mu})^{-K}(1+\frac{\Lambda}{\mu})$, and $\frac{1}{1-\alpha(K)} \geq \underline{\omega}(K)$ for all $K \geq 1$ by comparing (24) and the definition of $\underline{\omega}(K)$ via induction. This implies that $\frac{1}{1-\alpha(K)} \geq (K!)(\frac{\Lambda}{\mu})^{-K}(1+\frac{\Lambda}{\mu})$, and hence

$$\limsup_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} \leq \frac{1}{1+\frac{\Lambda}{\mu}} < 1. \tag{26}$$

On the other hand, for any $k > \frac{\Lambda}{\mu}$, let $\bar{\omega}^k(K) := K(\frac{\Lambda}{\mu})^{-1}\bar{\omega}^k(K-1)$ be recursively defined for all $K \geq k$, where $\bar{\omega}^k(k-1) := \frac{1}{1-\alpha(k-1)} + \frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}$. Then, $\bar{\omega}^k(K) = (K!)(\frac{\Lambda}{\mu})^{-K} \cdot \frac{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}}{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}$, and $\bar{\omega}^k(K) \geq \frac{1}{1-\alpha(K)} + \frac{\frac{\Lambda}{\mu}}{K-\frac{\Lambda}{\mu}}$ for all $K \geq k$ by comparing (25) and the definition of $\bar{\omega}^k(K)$ via induction. This implies that

$$\liminf_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} \geq \liminf_{K\to\infty}\frac{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}}\left[1+\frac{\frac{\Lambda}{\mu}(1-\alpha(K))}{K-\frac{\Lambda}{\mu}}\right] = \frac{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}} > 0. \tag{27}$$

Let $k \to \infty$, we obtain

$$\liminf_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} \geq \limsup_{k\to\infty}\frac{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}} \geq \frac{\limsup_{k\to\infty}\left\{[1-\alpha(k-1)](k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}\right\}}{\liminf_{k\to\infty}\left\{1+\frac{\frac{\Lambda}{\mu}[1-\alpha(k-1)]}{k-\frac{\Lambda}{\mu}}\right\}}$$
$$= \limsup_{k\to\infty}\left\{[1-\alpha(k)](k!)(\frac{\Lambda}{\mu})^{-k}\right\},$$

which suggests that

$$\liminf_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} = \limsup_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\}$$

and $\lim_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\}$ exits. Its value lies in $(0,1)$ due to (26) and (27). □

**Proof of Lemma 4.** The proofs of properties 1 and 2 are straightforward. Property 1 obviously holds because the recursive equation (8) depends only on the ratio of $\Lambda$ to $\mu$. From (8), we can also easily prove, by induction, that $\alpha(K,\Lambda,\mu,N)$ decreases in $N$, which immediately yields property 2.

If (8) is extended to be appropriately defined for real-valued $K$, then $K(\alpha)$ can be re-defined as the continuous inverse of $\alpha(K)$ such that $K(\alpha) := \{K : \alpha(K) = \alpha\}$. In this case, (9) holds for real-valued $K$, which yields

$$K(\alpha) = \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\Delta\alpha\big(K(\alpha)\big).$$

By Lemma 3 and the proof of result 2 of Proposition 3, we obtain

$$1 - \alpha^s = \lim_{\Lambda\to\infty}\frac{\Lambda}{\mu}\eta_s^0(L_0, U_0, \alpha) < \frac{\Lambda}{\mu}\Delta\alpha\big(K(\alpha)\big) < \lim_{\Lambda\to\infty}\frac{\Lambda}{\mu}\zeta_s^1(L_0, U_0, \alpha) = 1,$$

which, by letting $s \to \infty$ as $\Lambda \to \infty$, proves that $\lim_{\Lambda\to\infty}\{K(\alpha, \Lambda, \mu, N) - (\frac{\Lambda}{\mu}\alpha + N\frac{\alpha}{1-\alpha})\} = 0$. This result is sufficient to guarantee property 4.

The proof of property 3 is more involved. Note that the recursive equation (8) depends on the $\lambda_i$'s only via $\Lambda$ and the $\mu_{ij}$'s only via $\mu$. Therefore, balanced systems with the same total arrival rate and overall average rental time have the same performance even though they may have different $\lambda_i$'s, $\mu_{ij}$'s, and rental time distributions. This implies that it is sufficient to prove that property 3 holds for a symmetric system where each region has the same demand rate $\lambda$ and each customer has the same exponentially distributed rental time with mean $\frac{1}{\mu}$. We may also further restrict the routing matrix $\boldsymbol{P}$ as follows:

$$p_{ij} = \begin{cases} 1, & \text{if } j = i+1, \\ 1, & \text{if } i = N, j = 1, \\ 0, & \text{otherwise,} \end{cases}$$

which implies that a vehicle picked up at location $i$ is dropped off at location $i+1$ when $i < N$ and vehicles picked up at location $N$ are dropped off at location 1. A graphic representation of this *cyclic* network is shown in Figure 4. We use the notation $X_i$ and $Y_i$ to denote the number of vehicles at location $i$'s pick-up queue and transit queue respectively. In this network, the throughput of all the queues are the same and we use the notation $f_N(K)$ to denote the throughput rate per location when the cyclic network has $K$ vehicles and $N$ locations. This throughput is a concave function of $K$ for any $N$ (a result we use below) since the total throughput of a closed Jackson queueing network is nondecreasing concave in the number of items in the network (Shanthikumar and Yao (1988)).

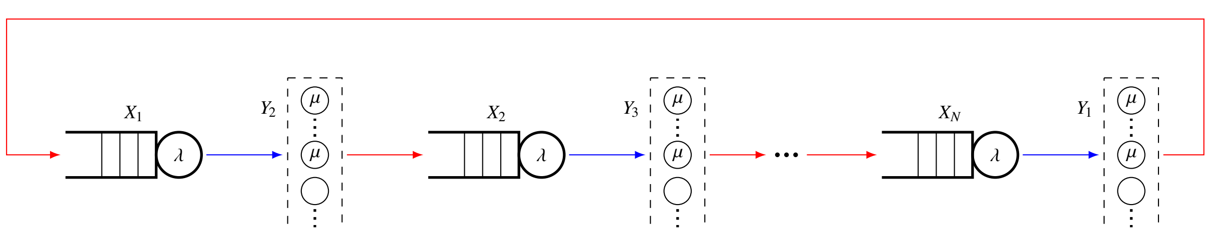



Figure 4: A cyclic queueing network representation of a one-way vehicle sharing system

Next, we describe a network aggregation procedure due to Chandy et al. (1975) that we will deploy in our proof. Consider a closed queuing network with exponential service queues. A subnetwork $\sigma$ is a subset of queues in the network such that items enter this subset through only one starting point and exit it through only one endpoint. For example, the transit queue 2 and pick-up queue 2 in Figure 4 constitute a subnetwork, while the transit queues 2 and 3 do not. A reduced network with $\sigma$ "shorted" is a modification to the original network in which the service times of all the servers in the subnetwork $\sigma$ are set to zero. Let $T(K)$ be the throughput rate passing the endpoint of the shorted subnetwork when there are a total number of $K$ items in the reduced network. Then, we construct an equivalent network with a composite $\sigma^c$ by replacing all the queues in the original network, except those in the subnetwork $\sigma$, by a single composite queue which has a state dependent service rate $T(k)$ when there are $k$ items in its queue. That is, the equivalent network consists of the queues in the original subnetwork $\sigma$ and a single composite queue.

**Lemma 7** (Theorems 1 & 2 of Chandy et al. (1975))**.** *The behavior of $\sigma$ in the equivalent network is identical to that in the original network, i.e., they have the same queue length and queue time distributions.*

For our cyclic network with $2N$ queues and $NK$ items (see Figure 4), we propose the following aggregation procedure, consistent with the procedure described above, to construct an equivalent network with $N$ indentical queues and $NK$ items. Specifically, we sequentially aggregate, starting with location 1, the pickup and transit queues into a single queue (see Figure 5 for an illustration). The resulting equivalent network (with the same throughput per location) has $N$ identical queues, with each queue having a queue-length dependent service rate function $f_1(\cdot)$. Let $Z_i, i = 1, 2, \cdots, N$ denote the length of each queue in the equivalent network. By definition, the total throughput of the equivalent network can be derived as

$$\sum_{Z_1+Z_2+\cdots+Z_N=NK} [P(Z_1, Z_2, \cdots, Z_N) \sum_{i=1}^{N} f_1(Z_i)],$$

which equals the total throughput rate of all pick-up queues in the original network, yielding

$$\begin{aligned} Nf_N(NK) =& \sum_{Z_1+Z_2+\cdots+Z_N=NK} [P(Z_1, Z_2, \cdots, Z_N) \sum_{i=1}^{N} f_1(Z_i)] \\ \leq& \sum_{Z_1+Z_2+\cdots+Z_N=NK} [P(Z_1, Z_2, \cdots, Z_N) N f_1(K)] \\ =& N f_1(K), \end{aligned}$$

where the inequality holds by the concavity of $f_1(\cdot)$ and Jensen's inequality. Noting $\alpha(NK, N\lambda, \mu, N) = \frac{f_N(NK)}{\lambda}$ and $\alpha(K, \lambda, \mu, 1) = \frac{f_1(K)}{\lambda}$, we have $\alpha(NK, N\lambda, \mu, N) \leq \alpha(K, \lambda, \mu, 1)$. □

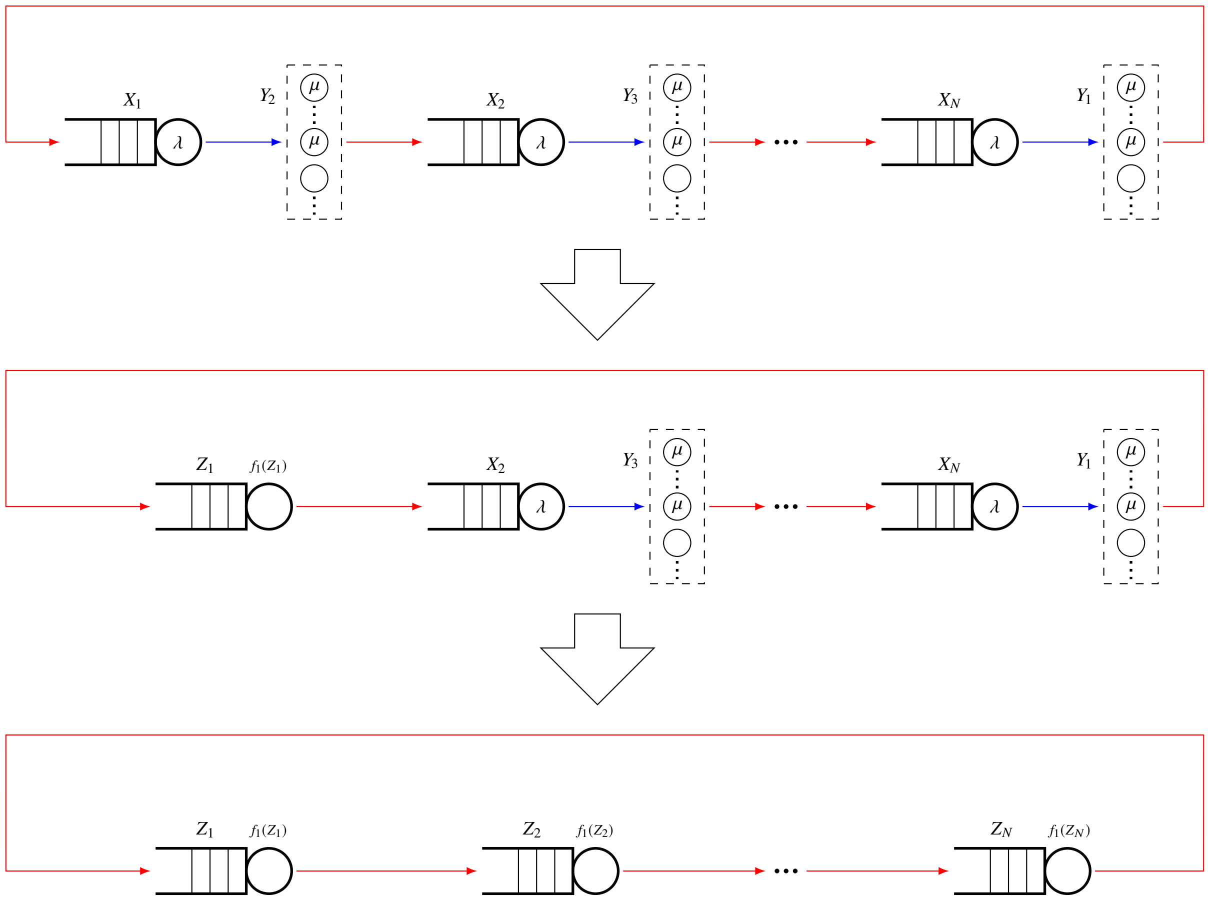



Figure 5: A graphical illustration of the aggregation procedure

## Appendix B Dimensioning Unbalanced Networks

In this section, we derive recursive equations for unbalanced networks that are analogous to (8) for balanced networks. These equations allow us to efficiently compute the average system performance and determine the optimal fleet size, though not yielding closed-form expressions. We discuss how in an unbalanced network, the problem of determining a fleet size that guarantees a specified service level at each location may not have a feasible solution. That is, even with an infinitely large number of vehicles, it may not be possible to achieve a target service level (if this target is sufficiently high) at each location; these results are due to George and Xia (2011) and we refer the interested reader to their paper for more details.

First, note that the network continues to be a BCMP Network. Therefore, the stationary distribution of system states can still be obtained via (1)–(3) and used to compute various performance measures, including throughput at each transit and pickup queue, total system throughput, and service level at each location. Note that because $\lambda_i = \sum_j \lambda_j p_{ji}$ does not hold for all $i$, the service level induced in steady state by a given number of vehicles is no longer the same at different locations. Hence, the dimensioning problem becomes one of finding the smallest number of vehicles that guarantees that the smallest service level is greater or equal than the specified target service level. As in the case of a balanced network, this approach requires significant computational effort and lacks interpretability.

Mean value analysis can alternatively be used. This requires some modifications from the balanced system case which we describe next. Let $r_{ij}$ be the proportion of effective rentals that originate in location $i$ and terminate in location $j$ when there is only a single vehicle in the system (i.e., $K = 1$) and let $r_i = \sum_{j \in V} r_{ij}$. From the point of view of this vehicle, its transitions between locations, regardless of the lengths of stay, are governed by the transition probabilities $p_{ij}$. Therefore, the $r_i$'s are the steady-state probabilities of a discrete time Markov Chain with transition matrix with elements $\{p_{ij}\}$ (which can be easily computed by substituting $\nu_i$ with $r_i$ in (3) and requiring $\sum_{i \in V} r_i = 1$), and $r_{ij} = r_i p_{ij}$. Noting that the above argument does not involve the lengths of stay which depend on the demand rates and the number of vehicles, it applies to each individual vehicle even when there are multiple vehicles in the system. Therefore, the proportion of effective rentals that originate in location $i$ and terminate in location $j$ in a system with $K$ vehicles always equals $r_{ij}$ regardless of $K$, i.e., $\frac{\nu_{ij}(K)}{\nu(K)} = r_{ij}$. From the balance equation of the above-mentioned single-vehicle Markov Chain, we know that $r_i = \sum_{j \in V} r_{ij} = \sum_{j \in V} r_{ji}$. That is, $r_i$ is the proportion of effective rentals that originate in location $i$ and also the proportion of effective rentals that terminate in location $i$.

Noting that equations (4) and (5) also hold for unbalanced networks and that the first equality of (6) remains valid leads to the following modified version of (6)

$$\mathbb{E}[X_i(K)] = \nu_i(K)\frac{1 + \mathbb{E}[X_i(K-1)]}{\lambda_i} = \frac{r_i}{\lambda_i}\big(1 + \mathbb{E}[X_i(K-1)]\big)\nu(K), \tag{28}$$

and

$$\sum_i \mathbb{E}[X_i(K)] = \nu(K) \sum_i \Big\{\frac{r_i}{\lambda_i}(1 + \mathbb{E}[X_i(K-1)])\Big\}. \tag{29}$$

Substituting (5) into (29) leads to

$$\nu(K) = \frac{K\lambda\mu}{\lambda + \mu + \lambda\mu \sum_{i \in V} \big(\frac{r_i}{\lambda_i}\mathbb{E}[X_i(K-1)]\big)}, \tag{30}$$

where $\lambda := (\sum_{i \in V} \frac{r_i}{\lambda_i})^{-1}$ and $\mu = \big(\sum_{i,j \in V} \frac{r_{ij}}{\mu_{ij}}\big)^{-1}$. By letting $\mathbb{E}[X_i(0)] = 0$, equations (30) and (28) hold for $K \geq 1$, and can be used recursively to compute the throughput rate $\nu(K)$ and to do so efficiently (note that, as with a balanced network, the computational effort does not depend on the state space). Having obtained $\nu(K)$, other performance measures can be derived, including $\nu_i(K) = r_i\nu(K)$, $\nu_{ij}(K) = r_{ij}\nu(K)$, and $\alpha_i(K) = \frac{\nu_i(K)}{\lambda_i}$ ($\alpha_i(K)$ can also be rewritten as $\alpha_i(K) = \frac{r_i}{\lambda_i}\nu(K)$, where $\frac{r_i}{\lambda_i}$ is independent of $K$), which implies that $\alpha_i(K) > \alpha_j(K)$ if an only if $\frac{r_i}{\lambda_i} > \frac{r_j}{\lambda_j}$.

Let $B = \{i \in V : \frac{r_i}{\lambda_i} \geq \frac{r_j}{\lambda_j}, \forall j \in V\}$ denote the set of locations with the largest ratio $\frac{r_i}{\lambda_i}$ (because the system is unbalanced, not all the ratios can be equal and the set $B$ is a proper subset of $V$). Let $\alpha_i(\infty) := \lim_{K\to\infty} \alpha_i(K)$ and define similarly $\nu_i(\infty)$, $\mathbb{E}[X_i(\infty)]$, and $\mathbb{E}[Y_{ij}(\infty)]$. The following proposition describes several useful properties of unbalanced systems, including asymptotic results as $K$ becomes large.

**Proposition 7** (Reproduction of Theorems 1 of George and Xia (2011) with Minor Extension)**.** [7]

*(i).* $\alpha_i(K)$, $\nu_i(K)$, $\mathbb{E}[X_i(K)]$, *and* $\mathbb{E}[Y_{ij}(K)]$ *increase in* $K$, *for all* $i, j \in V$.

*(ii).* $\alpha_i(\infty) = 1$ *and* $\alpha_j(\infty) = \frac{r_j\lambda_i}{\lambda_j r_i} < 1$, *for all* $i \in B,\ j \in V \setminus B$.

*(iii).* $\nu_i(\infty) = \lambda_i$ *and* $\nu_j(\infty) = \frac{r_j}{r_i}\lambda_i < \lambda_j$, *for all* $i \in B,\ j \in V \setminus B$.

*(iv).* $\mathbb{E}[X_i(\infty)] = \infty$ *and* $\mathbb{E}[X_j(\infty)] = \frac{\alpha_j(\infty)}{1-\alpha_j(\infty)} < \infty$, *for all* $i \in B,\ j \in V \setminus B$.

*(v).* $\mathbb{E}[Y_{ij}(\infty)] = \frac{\lambda_i p_{ij}}{\mu_{ij}}\alpha_i(\infty)$, *for all* $i, j \in V$.

**Proof of Proposition 7.** We prove results (i)–(v) in sequence.

(i). It is easy to prove by the coupling technique that $\nu(K)$, $\mathbb{E}[X_i(K)]$, and $\mathbb{E}[Y_{ij}(K)]$ increase in $K$. Because $\nu_i(K) = r_i\nu(K)$ and $\alpha_i(K) = \frac{\nu_i(K)}{\lambda_i}$, they increase in $K$ as well. Since $\alpha_i(K)$ is bounded above by 1, the limit $\lim_{K\to\infty} \alpha_i(K)$ exists.

Moreover, by the expression $\alpha_i(K) = \frac{r_i}{\lambda_i}\nu(K)$, we know that $\frac{\alpha_j(K)}{\alpha_i(K)} = \frac{r_j\lambda_i}{\lambda_j r_i}$, $\forall i, j \in V,\ K \geq 1$, and hence $\alpha_i(K) > \alpha_j(K)$ if an only if $\frac{r_i}{\lambda_i} > \frac{r_j}{\lambda_j}$.

(ii). Because $\nu(K)$ is bounded above by $\Lambda$, we know by (30) that $\lim_{K\to\infty} \sum_{i\in V} \left(\frac{r_i}{\lambda_i}\mathbb{E}[X_i(K-1)]\right) = \infty$. There must exist at least one $i' \in V$ such that $\lim_{K\to\infty} \mathbb{E}[X_{i'}(K-1)] = \infty$. Noting that equation (28) can be written as

$$\mathbb{E}[X_i(K)] = \alpha_i(K)\big(1 + \mathbb{E}[X_i(K-1)]\big), \tag{31}$$

we must have $\alpha_{i'}(\infty) = 1$ since it will lead to a contradiction otherwise. According to result (i), we know that this $i'$ must be in $B$. Otherwise, $\alpha_i(\infty)$ is greater than 1 for $i \in B$. Thus, we have $\alpha_i(\infty) = 1$, and hence $\alpha_j(\infty) = \frac{r_j\lambda_i}{\lambda_j r_i} < 1$, $\forall i \in B,\ j \in V \setminus B$.

(iii). This result follows directly from result (ii) and the expression $\nu_i(K) = \alpha_i(K)\lambda_i$.

(iv). We prove by contradiction. Since $\mathbb{E}[X_i(K)]$ increases in $K$, it either grows unboundedly or converges to a finite value as $K \to \infty$. If $\lim_{K\to\infty} \mathbb{E}[X_i(K)] = C < \infty$ for some $i \in B$, then by letting $K \to \infty$ on both sides of (31) we have $C = 1 + C$, leading to a contradiction. Thus, $\lim_{K\to\infty} \mathbb{E}[X_i(K)] = \infty$ for any $i \in B$.

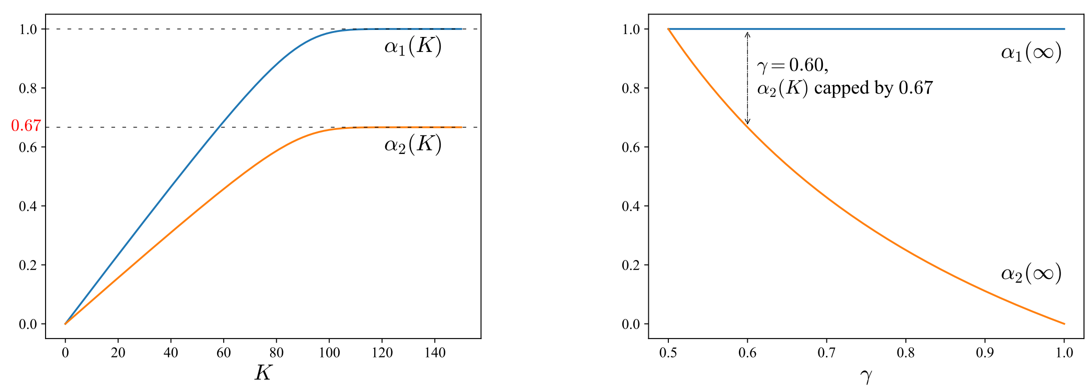
Because $\lim_{K\to\infty} \alpha_j(K) < 1$ for $j \in V \setminus B$ and $\mathbb{E}[X_j(0)] = 0$, it is also straightforward to see from (31) that $\lim_{K\to\infty} \mathbb{E}[X_j(K)] < \infty$. By letting $K \to \infty$ on both sides of (31), we have $\lim_{K\to\infty} \mathbb{E}[X_j(K)] = \frac{\alpha_j(\infty)}{1-\alpha_j(\infty)} < \infty$, $\forall j \in V \setminus B$.

[7] The proposition recasts Theorem 1 of George and Xia (2011) using our notation and it extends by including $\nu_i(\infty)$ and $\mathbb{E}[X_i(\infty)]$.

(v). By Little's Law, $\mathbb{E}[Y_{ij}(K)] = \frac{\nu_{ij}(K)}{\mu_{ij}} = \frac{\nu_i(K)p_{ij}}{\mu_{ij}} = \frac{\alpha_i(K)\lambda_i p_{ij}}{\mu_{ij}}$.

□

Proposition 7 shows while some location(s) can have arbitrarily high service level as the total number of vehicles increases, the service levels at other locations are bounded by specific thresholds. That is, it is impossible to achieve service levels above these thresholds even with an infinite number of vehicles. The average numbers of vehicles at locations in $V \setminus B$ and in transit are also bounded by finite fixed thresholds (no matter how large is the total number of vehicles) and so are the associated throughputs. These results are illustrated for an example with two locations in Figure 6. Figure 6 shows how even modest differences in the relative popularity of locations (either as origins or destinations) can lead to significant differences in the achievable service levels at these locations.




(a) $\gamma = 0.6$

(b) $K = \infty$

Figure 6: A two-location unbalanced network with $\lambda_1 = \lambda_2 = 50$, $\mu_{ij} = 1$, $p_{11} = p_{21} = \gamma$, and $p_{12} = p_{22} = 1 - \gamma$.

The intuition for the results is as follows. Note that in an unbalanced network, locations may be more popular as an origin or as a destination. The popularity of a location as a destination relative to it being an origin is indicated by the ratio $\frac{r_i}{\lambda_i}$, which is referred to as the relative utilization of the location by George and Xia (2011). In the long run, vehicles accumulate at locations that are most popular as destination (relative to their popularity as origin), i.e., with the greatest $\frac{r_i}{\lambda_i}$. These locations are referred to as bottleneck locations of the closed network. In this case, the supply of vehicles the non-bottleneck locations receive depends on the bottleneck throughput rates, which are bounded by the demand rates at the bottleneck locations regardless of the fleet size. Therefore, no matter how large the fleet size is and how many vehicles are initially provisioned to non-bottleneck locations, a large (majority) number of the vehicles will later accumulate and stay idle at the bottleneck locations, guaranteeing an arbitrarily high service level there and at the same time making the service levels at non-bottleneck locations bounded by specific thresholds in the long run.

## Appendix C An Application: Optimizing the Service Level

In this section, we briefly illustrate how the minimal fleet size approximation in (15) can be embedded in an optimization problem. In particular, we illustrate how the service level can be endogenized by letting it to be a decision that the service provider makes. We do not intend this to be a full treatment of the problem, but simply an illustration of the usefulness of an approximation in supporting operational decision making and in obtaining additional managerial insights[8].

Let $r$ denote the price the service provider charges for each rental per unit time the vehicle is rented. Let also $c$ denote the cost of a vehicle per unit of time (this cost may include the amortized purchase cost of the vehicle as well as its operating cost). Assume $c < r$ (otherwise offering the service would not be profitable) and assume that the network is balanced. The service provider's profit maximization problem can be stated as follows:

$$\begin{aligned}\max_\alpha \pi(\alpha) &= \max_\alpha \left\{ r\Lambda\frac{1}{\mu}\alpha - c\hat{K}(\alpha,\Lambda,\mu,N)\right\} \\ &= \max_\alpha \left\{ \frac{r\Lambda\alpha}{\mu} - c\left[\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\frac{N\mu}{1-\alpha} + \Lambda(1-\alpha)}\alpha\right]\right\}. \end{aligned} \tag{32}$$

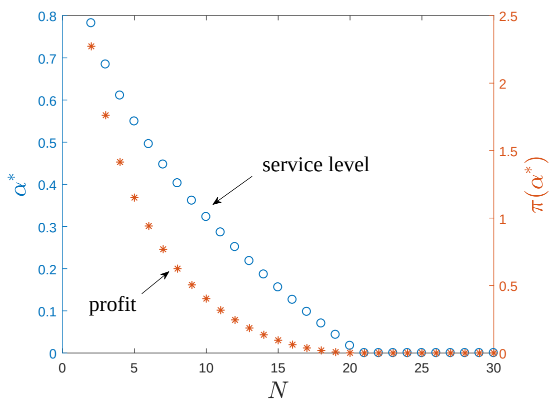
**Proposition 8.** *$\pi(\alpha)$ is concave in $\alpha$. The optimal service level, $\alpha^*(N,\Lambda)$, is unique and has the following properties:*

1. *$\alpha^*(N,\Lambda) > 0$ if and only if $c \leq \frac{\frac{r\Lambda}{\mu}}{\frac{\Lambda}{\mu}+N-1+\frac{\Lambda}{N\mu+\Lambda}}$;*
2. *for fixed $\Lambda$, $\alpha^*(N,\Lambda)$ decreases in $N$ and $\lim\limits_{N\to\infty} \alpha^*(N,\Lambda) = 0$;*
3. *for fixed $N$, $\alpha^*(N,\Lambda)$ increases in $\Lambda$ and $\lim\limits_{\Lambda\to\infty} \alpha^*(N,\Lambda) = 1$; and*
4. *for $\frac{\Lambda}{N} \equiv \lambda$, $\alpha^*(N,\Lambda)$ decreases in $N$. If $\lambda < \frac{c\mu}{r-c}$, then $\alpha^*(N,\Lambda) > 0$ holds only for $N \leq \frac{c\mu^2}{(\mu+\lambda)(r-c)(\frac{c\mu}{r-c}-\lambda)}$; otherwise, $\alpha^*(N,\Lambda) > 0$ holds for any $N$ and $\lim\limits_{N\to\infty} \alpha^*(N,\Lambda) = 1 - \sqrt{\frac{c\mu}{(r-c)\lambda}}$.*

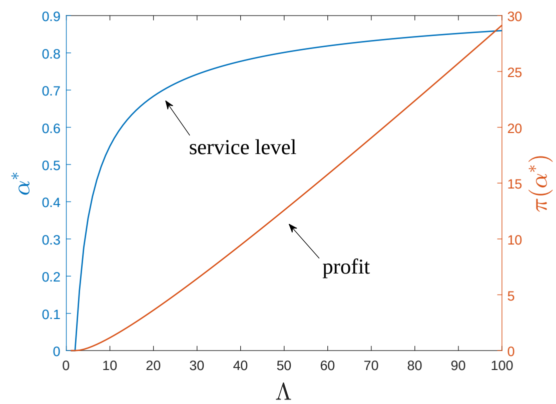
The proposition shows that the problem is concave and, hence, admits a unique solution. Property 1 in the proposition provides a necessary and sufficient condition for the service provider to realize a positive profit. Property 2 shows that, all else being equal, there is a tradeoff between location density (number of locations) and service level. Property 3 shows that, perhaps surprisingly, increased demand does not lead to a deterioration in the service level but instead to an increase (because of the pooling effect, the service provider makes additional investments in vehicles resulting in a higher service level). Property 4 shows that there is a tradeoff between the size of the service region (where an increase in demand requires an increase in the number of locations) and service level. If the demand density (average demand per location) is sufficiently

[8] It is possible to consider other problems, including determining location density (the optimal number of locations for fixed demand), sizing of the service region (optimal demand level), and service pricing (optimal price to charge).

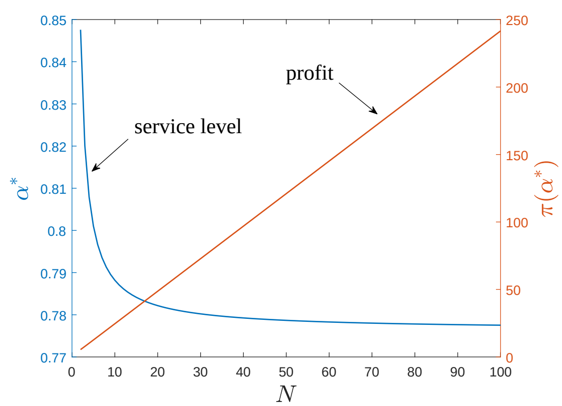
large, then the firm is always profitable no matter how large the service region is; otherwise, the firm is profitable only when the number of locations is sufficiently small. The underlying reason why the limit exists is that, as $N \longrightarrow \infty$, the correction term $B_0(1 - \frac{1}{N})$ approaches the standard buffer $B_0$ and the minimal fleet size reduces to $\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha}$. That is, buffer capacity $(N-1)\frac{\alpha}{1-\alpha}$ is sufficient to protect against both vehicle roaming and randomness in demand and service times (see also the discussion in Section 6). These results are illustrated graphically in Figure 7.




(a) $\Lambda = 10$




(b) $N = 5$




(c) $\Lambda/N = 10$

Figure 7: The optimal service level ($c = 0.2$, $r = 0.6$, $\mu = 1$)

**Proof of Proposition 8.** Since

$$\pi''(\alpha) = -2c\frac{(2N-1)(\frac{\Lambda}{\mu})^2(1-\alpha)^2 + 2(N-1)\frac{N^2}{(1-\alpha)^2} + N\frac{\Lambda}{\mu}(4N-3-3\alpha)}{[\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)]^2(1-\alpha)^3} < 0,$$

$\pi(\alpha)$ is concave in $\alpha$ and the profit maximization problem has a unique solution $\alpha^*(N,\Lambda)$. Furthermore, the first order condition

$$\pi'(\alpha^*) = r\frac{\Lambda}{\mu} - c\Big[\frac{\Lambda}{\mu} + (N-1)\frac{1}{(1-\alpha^*)^2} + \frac{(\frac{\Lambda}{\mu})^2 + \frac{N\Lambda(1-2\alpha^*)}{\mu(1-\alpha^*)^2}}{[\frac{N}{1-\alpha^*} + \frac{\Lambda}{\mu}(1-\alpha^*)]^2}\Big] = 0 \tag{33}$$

is satisfied if $\alpha^*(N,\Lambda)$ is in $(0,1)$. Next, we prove properties 1–4.

(1). Noting that the concavity of $\pi(\alpha)$ and that $\pi(0) = 0$, the solution $\alpha^*(N,\Lambda)$ is non-zero if and only if

$$\pi'(0) = r\frac{\Lambda}{\mu} - c[\frac{\Lambda}{\mu} + (N-1) + \frac{\Lambda}{N\mu + \Lambda}] > 0. \tag{34}$$

(2). Since

$$\frac{\partial^2 \pi}{\partial\alpha\partial N}(\alpha) = -c\frac{N^3 + (\frac{\Lambda}{\mu})^3(1-\alpha)^6 + (\frac{\Lambda}{\mu})^2(1-\alpha)^4(3N-1-2\alpha) + N(\frac{\Lambda}{\mu})(1-\alpha)^2(3N-1+2\alpha)}{(1-\alpha)^2[N + \frac{\Lambda}{\mu}(1-\alpha)^2]^3} \le 0,$$

the profit function is submodular in $(\alpha, N)$ and hence the optimal service level $\alpha^*(N,\Lambda)$ decreases in $N$. Moreover, we know by property 1 that $\alpha^*(N,\Lambda) = 0$ when $N$ is sufficiently large.

(3). Note that

$$\frac{\partial^2 \pi}{\partial\alpha\partial\Lambda}(\alpha) = r - c - c\frac{N\frac{\Lambda}{\mu}\frac{1+2\alpha}{1-\alpha} + N^2\frac{1-2\alpha}{(1-\alpha)^3}}{[\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)]^3}.$$

By property 1, $\alpha^*(N,\Lambda) > 0$ and hence the first order condition (33) holds at $\alpha^*(N,\Lambda)$ when $\Lambda$ is sufficiently large. In this case, by substituting (33) into the above expression, we know

$$\begin{aligned}\frac{\partial^2 \pi}{\partial\alpha\partial\Lambda}(\alpha^*) =& \frac{c\mu}{\Lambda(1-\alpha^*)^2}\frac{(\frac{\Lambda}{\mu})^2 N(3N-2-4\alpha^*)(1-\alpha^*) + N^3(N-1)\frac{1}{(1-\alpha^*)^2} + (\frac{\Lambda}{\mu})^3 N(1-\alpha^*)^3 + 3\frac{\Lambda}{\mu}N^2(N-1)\frac{1}{1-\alpha^*}}{[\frac{N}{1-\alpha^*} + \frac{\Lambda}{\mu}(1-\alpha^*)]^3}\\ \ge& 0,\end{aligned}$$

which implies that $\alpha^*(N,\Lambda)$ increases in $\Lambda$ and $\lim_{\Lambda\to\infty}\alpha^*(N,\Lambda)$ exists by the monotone convergence theorem. If $\bar{\alpha} = \lim_{\Lambda\to\infty}\alpha^*(N,\Lambda) < 1$, then (33) fails when $\Lambda$ is sufficiently large, which causes a contradiction. Therefore, $\bar{\alpha} = \lim_{\Lambda\to\infty}\alpha^*(N,\Lambda) = 1$.

(4). When $\frac{\Lambda}{N} = \lambda$ is held fixed, the first order condition (33) and the partial derivative $\frac{\partial^2\pi}{\partial\alpha\partial N}(\alpha; N, \lambda)$ can be rewritten as

$$\pi'(\alpha^*; N, \lambda) = r\frac{N\lambda}{\mu} - c\Big[\frac{N\lambda}{\mu} + (N-1)\frac{1}{(1-\alpha^*)^2} + \frac{(\frac{\lambda}{\mu})^2 + \frac{\lambda(1-2\alpha^*)}{\mu(1-\alpha^*)^2}}{[\frac{1}{1-\alpha^*} + \frac{\lambda}{\mu}(1-\alpha^*)]^2}\Big] = 0 \tag{35}$$

and

$$\frac{\partial^2 \pi}{\partial \alpha \partial N}(\alpha; N, \lambda) = r\frac{\lambda}{\mu} - c\frac{\lambda}{\mu} - c\frac{1}{(1-\alpha)^2}.$$

By substituting (35) into $\frac{\partial^2 \pi}{\partial \alpha \partial N}(\alpha; N, \lambda)$, we know

$$\frac{\partial^2 \pi}{\partial \alpha \partial N}(\alpha^*; N, \lambda) = -\frac{c}{N}\frac{\frac{\lambda}{\mu}(1+2\alpha) + \frac{1}{(1-\alpha)^2}}{[\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)]^2(1-\alpha)^2} \leq 0.$$

Therefore, $\alpha^*(N, \Lambda)$ decreases in $N$ when $\frac{\Lambda}{N}$ is held fixed.

Given $\frac{\Lambda}{N} = \lambda$, the firm earns a positive profit if and only if (by rewriting property 1)

$$N(r\frac{\lambda}{\mu} - c\frac{\lambda}{\mu} - c) > -\frac{c\mu}{\mu + \lambda},$$

whick holds for $N \leq \frac{c\mu^2}{(\mu+\lambda)(r-c)(\frac{c\mu}{r-c}-\lambda)}$ if $\lambda < \frac{c\mu}{r-c}$ and for all $N > 0$ otherwise. In the latter case, $\lim\limits_{N\to\infty} \alpha^*(N, \Lambda; \frac{\Lambda}{N} = \lambda)$ exists by the monotone convergence theorem. Let $N \to \infty$ in equality (35), we obtain $\lim\limits_{N\to\infty} \alpha^*(N, N\lambda) = 1 - \sqrt{\frac{c\mu}{(r-c)\lambda}}$.

$\square$

# Appendix D Additional Discussion of the Single Location Case

## D.1 A Correction Term

Recall that per Proposition 6 our approximation has a minor shortcoming when $N = 1$ since it does not approach infinity as $\alpha \to 1$. Thus, a question that naturally arises is whether we should add a correction term (a fourth term) to our approximation (15) for $N = 1$ to ensure that the approximation is consistent with the fact that the minimal fleet size grows to infinity as $\alpha$ approaches 1. If we do so, this correction term would be noticeable only when $N = 1$ and $\alpha$ is nearly 1. Note that such a term cannot be expressed in algebraic form because it grows slower than $\frac{1}{(1-\alpha)^{\delta}}$ for any $\delta > 0$ as $\alpha \to 1$ per Proposition 6. In most practical cases, adding this correction term would not be necessary. Extensive numerical experiments confirm that the correction would be small for $\alpha \leq 0.99$ no matter how large $\Lambda$ is. The largest value for the difference $|K(\alpha, \Lambda, \mu, 1) - \hat{K}(\alpha, \Lambda, \mu, 1)|$ is 31.97, which is observed to occur when $\frac{\Lambda}{\mu} = 1826$ and $\alpha = 0.99$. In this case, $K(\alpha, \Lambda, \mu, 1)$ is very large so that the percentage error is rather negligible. If we were to restrict our attention to $\alpha \leq 0.95$, the largest gap reduces to 6.90.

It is important to note that most of the existing approximations in the literature share a similar limitation. In particular, neither (H.39) nor (H.40) increases to infinity as $\alpha \to 1$ (i.e., even the best approximations in the literature suffer from this relatively minor drawback).

If a correction term must be included, we propose adding the following term:

$$\kappa(\alpha, \Lambda, \mu, N) = \ln(1 + \frac{\Lambda}{\mu}) \ln(1 + \alpha) \ln\left(\frac{\frac{N}{(1-\alpha)^2} + \frac{\Lambda}{\mu}}{\frac{N-1}{(1-\alpha)^2} + \frac{\Lambda}{\mu}}\right),$$

resulting in a modified expression for $\hat{K}(\alpha, \Lambda, \mu, N)$ given by:

$$\hat{K}(\alpha, \Lambda, \mu, N) = \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\frac{\Lambda}{\mu}\alpha}{\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)} + \kappa(\alpha, \Lambda, \mu, N). \tag{36}$$

Note that $\kappa(\alpha, \Lambda, \mu, N)$ satisfies the property that it grows at a smaller rate than that of any power series $O((1-\alpha)^{-\delta})$ for $\delta > 0$. It also satisfies the property $\kappa(\alpha, \Lambda, \mu, N) \to 0$ under the five regimes considered in Proposition 3 and $\kappa(\alpha, \Lambda, \mu, 1) \to \infty$ as $\alpha \to 1$.

For the same numerical example considered in Section 5.2 (i.e., $N = 1$, $\mu = 1$, $\Lambda = 1i$, $i = 1, \cdots, 1000$, and $\alpha = 0.03j$, $j = 1, \cdots, 33$), we find that our modified approximation (36) (with a mean absolute error of 0.86 and a relative error of 1.02%) is comparable to the best one, (H.40) (with a mean absolute error of 0.64 and a relative error of 1.15%), from the literature.

## D.2 Comparisons with the Approximations of the Inverse Erlang loss formula in Berezner et al. (1998) and Harel (2010)

In this section, we list notable approximations of the inverse of the Erlang Loss formula derived by Berezner et al. (1998) and Harel (2010). For ease of reference, we rewrite these approximations using our notation, and add an initial "B" to equation numbers when we refer to expressions from Berezner et al. (1998) and an initial "H" when we refer to expressions from Harel (2010). Using the same numerical examples used by Berezner et al. (1998) and Harel (2010), we compare the performance of our approximations against theirs.

*Bounds from Berezner et al. (1998):*

$$K(\alpha, \Lambda, \mu, 1) < \frac{\Lambda}{\mu}\alpha + \frac{1}{1-\alpha}, \tag{B.5}$$

$$K(\alpha, \Lambda, \mu, 1) > \frac{\Lambda}{\mu}\alpha, \tag{B.6}$$

$$K(\alpha, \Lambda, \mu, 1) > \frac{\Lambda}{\mu}\alpha + (\frac{1}{1-\alpha} - 1) - \frac{3\mu}{\Lambda(1-\alpha)^3} - \alpha^{\frac{\Lambda}{\mu}\alpha}\left(\frac{2}{1-\alpha} + \frac{\Lambda}{\mu}\alpha\right), \textit{ and} \tag{B.21}$$

$$K(\alpha, \Lambda, \mu, 1) > \max\{(B.6), (B.21)\}. \tag{B.621}$$

*Bounds from Harel (2010):*

$$K(\alpha, \Lambda, \mu, 1) < \left(\frac{\Lambda}{2\mu} + \frac{1}{2(1-\alpha)}\sqrt{(\frac{\Lambda}{\mu})^2(1-\alpha)^2 + 4\frac{\Lambda}{\mu}(1-\alpha)}\right)\alpha, \quad K \geq 2, \Lambda > 0, \tag{H.35}$$

$$K(\alpha, \Lambda, \mu, 1) > \frac{\Lambda}{\mu} - \frac{1}{2} - \frac{3\Lambda}{2\mu}(1-\alpha) + \frac{\sqrt{4\frac{\Lambda}{\mu} + \left[\frac{\Lambda}{\mu}(1-\alpha) - 1\right]^2}}{2}, \quad K \geq 1, \Lambda > 0, \tag{H.36}$$

$$K(\alpha, \Lambda, \mu, 1) \approx \frac{\Lambda}{\mu}\alpha\frac{2 + \frac{\Lambda}{\mu}(1-\alpha)}{1 + \frac{\Lambda}{\mu}(1-\alpha)}, \textit{ and} \tag{H.39}$$

$$K(\alpha, \Lambda, \mu, 1) \approx \frac{\Lambda}{\mu} - 2\frac{\Lambda}{\mu}(1-\alpha) - 1 + \sqrt{(\frac{\Lambda}{\mu})^2(1-\alpha)^2 + 2\frac{\Lambda}{\mu} + 1}. \tag{H.40}$$

We first compare our approximations (15) and (36) against those in Berezner et al. (1998) using the same numerical example considered in their Table 1. Since Berezner et al. (1998) prove that their bounds are strict, they use a strict ceiling of (B.621) and a strict floor of (B.5) as the lower and upper bounds for $K(\alpha)$, respectively. The results in Table 2 show that our approximations (15) and (36) consistently perform better than (B.621) and (B.5).

Next, we provide comparisons against those in Harel (2010) using the same numerical example considered in their Tables 2, 3, and 4. Noting that Harel (2010) treats $K$ as a real value $K$, we follow their setting and keep decimal parts in our approximations. One can see that all the approximations (ours and theirs) perform

| $\alpha$ | $\frac{\Lambda}{\mu}$ | (B.621) | $K(\alpha)$ | (B.5) | $\lceil \hat{K} \rceil$ | $\lceil (36) \rceil$ |
|---|---|---|---|---|---|---|
| 0.99 | 1000 | 991 | 1029 | 1089 | 999 | 1011 |
| | 10000 | 9901 | 9970 | 9999 | 9950 | 9954 |
| | 100000 | 99070 | 99092 | 99099 | 99090 | 99091 |
| | 1000000 | 990097 | 990099 | 990099 | 990099 | 990099 |
| | 10000000 | 9900099 | 9900099 | 9900099 | 9900099 | 9900099 |
| 0.999 | 1000 | 1000 | 1072 | 1998 | 1000 | 1034 |
| | 10000 | 9991 | 10170 | 10989 | 10000 | 10030 |
| | 100000 | 99901 | 100293 | 100899 | 99991 | 100010 |
| | 1000000 | 999001 | 999697 | 999999 | 999500 | 999507 |
| | 10000000 | 9990700 | 9990925 | 9990999 | 9990909 | 9990910 |

Table 2: Comparisons with the approximations in Table 1 of Berezner et al. (1998)

well when $\alpha$ is not too close to 1. When $\alpha$ is small, all the approximations produce exact values. Significant gaps are observed for the approximations when $\alpha = 0.99$ and for some when $\alpha = 0.9$. We further look into the case of $\alpha = 0.99$ for different values of $\frac{\Lambda}{\mu}$. First, (H.35) significantly overestimates the exact value for $\alpha = 0.99$. When $\frac{\Lambda}{\mu} = 10$ and $\alpha = 0.99$, (H.39) performs the best. When $\frac{\Lambda}{\mu} = 100$ and $\alpha = 0.99$, our approximation (36) performs the best. When $\frac{\Lambda}{\mu} = 1000$ and $\alpha = 0.99$, (H.40) performs the best.

| $\alpha$ | (H.35) | Exact | (H.40) | (H.39) | (H.36) | $\hat{K}$ | (36) |
|---|---|---|---|---|---|---|---|
| 0.99 | 36.65 | 17.44 | 13.38 | 18.90 | 12.54 | 10.00 | 21.40 |
| 0.90 | 14.56 | 12.53 | 11.69 | 13.50 | 11.16 | 9.82 | 13.51 |
| 0.80 | 10.93 | 10.27 | 10.00 | 10.67 | 9.70 | 9.14 | 10.91 |
| 0.70 | 8.85 | 8.58 | 8.48 | 8.75 | 8.32 | 8.11 | 9.06 |
| 0.60 | 7.24 | 7.12 | 7.08 | 7.20 | 7.00 | 6.92 | 7.47 |
| 0.50 | 5.85 | 5.80 | 5.78 | 5.83 | 5.74 | 5.71 | 6.04 |
| 0.40 | 4.58 | 4.56 | 4.55 | 4.57 | 4.53 | 4.52 | 4.72 |
| 0.30 | 3.38 | 3.37 | 3.37 | 3.38 | 3.36 | 3.36 | 3.47 |
| 0.20 | 2.22 | 2.22 | 2.22 | 2.22 | 2.22 | 2.22 | 2.28 |
| 0.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.13 |
| 0.01 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |

Table 3: Comparisons with the results in Table 2 of Harel (2010) for $\frac{\Lambda}{\mu} = 10$

Hence, none of the approximations consistently outperforms the others when $\alpha = 0.99$. In general, (H.40) and (36) provide the best performance. While (H.40) and (15) appear to underestimate the exact values, (36) overestimates by a very small amount except for the case of $\alpha = 0.99$. In summary, all the

| $\alpha$ | (H.35) | Exact | (H.40) | (H.39) | (H.36) | $\hat{K}$ | (36) |
|---|---|---|---|---|---|---|---|
| 0.99 | 160.19 | 116.88 | 111.21 | 148.50 | 108.00 | 99.98 | 114.64 |
| 0.90 | 98.24 | 96.25 | 96.35 | 98.18 | 95.47 | 94.50 | 96.55 |
| 0.80 | 83.82 | 83.42 | 83.52 | 83.81 | 83.29 | 83.20 | 83.81 |
| 0.70 | 72.26 | 72.14 | 72.18 | 72.26 | 72.11 | 72.10 | 72.36 |
| 0.60 | 61.46 | 61.42 | 61.44 | 61.46 | 61.41 | 61.41 | 61.54 |
| 0.50 | 50.98 | 50.96 | 50.97 | 50.98 | 50.96 | 50.96 | 51.03 |
| 0.40 | 40.66 | 40.65 | 40.65 | 40.66 | 40.65 | 40.65 | 40.69 |
| 0.30 | 30.42 | 30.42 | 30.42 | 30.42 | 30.42 | 30.42 | 30.44 |
| 0.20 | 20.25 | 20.25 | 20.25 | 20.25 | 20.25 | 20.25 | 20.26 |
| 0.10 | 10.11 | 10.11 | 10.11 | 10.11 | 10.11 | 10.11 | 10.12 |
| 0.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 |

Table 4: Comparisons with the results in Table 3 of Harel (2010) for $\frac{\Lambda}{\mu} = 100$

| $\alpha$ | (H.35) | Exact | (H.40) | (H.39) | (H.36) | $\hat{K}$ | (36) |
|---|---|---|---|---|---|---|---|
| 0.99 | 1080.69 | 1028.85 | 1024.84 | 1080.00 | 1016.44 | 999.00 | 1010.40 |
| 0.90 | 908.91 | 908.32 | 908.55 | 908.91 | 908.24 | 908.18 | 908.60 |
| 0.80 | 803.98 | 803.91 | 803.94 | 803.98 | 803.90 | 803.90 | 804.00 |
| 0.70 | 702.33 | 702.31 | 702.32 | 702.33 | 702.31 | 702.31 | 702.35 |
| 0.60 | 601.50 | 601.49 | 601.49 | 601.50 | 601.49 | 601.49 | 601.51 |
| 0.50 | 501.00 | 501.00 | 501.00 | 501.00 | 501.00 | 501.00 | 501.01 |
| 0.40 | 400.67 | 400.66 | 400.67 | 400.67 | 400.66 | 400.66 | 400.67 |
| 0.30 | 300.43 | 300.43 | 300.43 | 300.43 | 300.43 | 300.43 | 300.43 |
| 0.20 | 200.25 | 200.25 | 200.25 | 200.25 | 200.25 | 200.25 | 200.25 |
| 0.10 | 100.11 | 100.11 | 100.11 | 100.11 | 100.11 | 100.11 | 100.11 |
| 0.01 | 10.01 | 10.01 | 10.01 | 10.01 | 10.01 | 10.01 | 10.01 |

Table 5: Comparisons with the results in Table 4 of Harel (2010) for $\frac{\Lambda}{\mu} = 1000$

approximations in Harel (2010) and our approximations (15) and (36) perform well as long as $\alpha$ is not too close to 1.

## References for Appendices


Chandy, K., U. Herzong, and L. Woo (1975). Parametric analysis of queueing networks. *IBM Journal of Research and Development 19*, 36.

Shanthikumar, J. G. and D. D. Yao (1988). Second-order properties of the throughput of a closed queueing network. *Mathematics of Operations Research 13*(3), 524–534.