APPENDIX

Proof of Lemma 1

For simplicity, we rewrite Equation (2) as two different equations:

$$u(n_\mathrm{a},n_\mathrm{t}) = \frac{-C}{\mu_\mathrm{t}+\mu_\mathrm{a}} + \frac{\mu_\mathrm{a}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(n_\mathrm{a}-1, n_\mathrm{t}+\mathbb{1}_{\{n_\mathrm{t}<N\}}) + \frac{\mu_\mathrm{t}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(n_\mathrm{a}, n_\mathrm{t}-\mathbb{1}_{\{n_\mathrm{t}>0\}}), \tag{10}$$

for $n_\mathrm{a}\geq 1$, and

$$u(n_\mathrm{a},n_\mathrm{t}) = \frac{-C}{\mu_\mathrm{t}+\mu_\mathrm{a}} + \frac{\mu_\mathrm{a}}{\mu_\mathrm{t}+\mu_\mathrm{a}} \left[R - \frac{n_\mathrm{t}+1}{\mu_\mathrm{t}}C\right]^+ + \frac{\mu_\mathrm{t}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(n_\mathrm{a}, n_\mathrm{t}-\mathbb{1}_{\{n_\mathrm{t}>0\}}), \tag{11}$$

for $n_\mathrm{a}=0$.

For state $(0,0)$, by substituting $n_\mathrm{t}=0$ in (11) we get

$$u(0,0) = \frac{-C}{\mu_\mathrm{t}+\mu_\mathrm{a}} + \frac{\mu_\mathrm{t}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(0,0) + \frac{\mu_\mathrm{a}}{\mu_\mathrm{t}+\mu_\mathrm{a}}\left(R - \frac{C}{\mu_\mathrm{t}}\right),$$

which, after rearranging gives:

$$u(0,0) = -\left(\frac{1}{\mu_\mathrm{a}} + \frac{1}{\mu_\mathrm{t}}\right)C + R.$$

This reflects the reward R minus the expected cost of time in service for a joining customer, which they undergo sequentially, with no delays, since the system is empty.

In state $(0,n_\mathrm{t}), n_\mathrm{t}\geq 1$, after joining, a customer starts service in admission, and waits in expectation $1/(\mu_\mathrm{t}+\mu_\mathrm{a})$ units of time until the first among the admission and the treatment server completes its current service. If service in the treatment queue finishes first, the expected remaining wait time is exactly the same as in joining the system at state $(0,n_\mathrm{t}-1)$. However, if the admission server finishes first, then the customer joins the treatment queue with n_t customers in it, unless $n_\mathrm{t}\geq N$, in which case they balk with no additional utility. Thus,

$$\begin{aligned} u(0,n_\mathrm{t}) &= -\frac{C}{\mu_\mathrm{t}+\mu_\mathrm{a}} + \frac{\mu_\mathrm{a}}{\mu_\mathrm{t}+\mu_\mathrm{a}}\left(-\frac{n_\mathrm{t}+1}{\mu_\mathrm{t}}C + R\right)\cdot \mathbb{1}_{\{n_\mathrm{t}<N\}} + \frac{\mu_\mathrm{t}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(0,n_\mathrm{t}-1) \\ &= -\frac{C}{\mu_\mathrm{t}+\mu_\mathrm{a}} + \frac{\mu_\mathrm{a}}{\mu_\mathrm{t}+\mu_\mathrm{a}}\left[-\frac{n_\mathrm{t}+1}{\mu_\mathrm{t}}C + R\right]^+ + \frac{\mu_\mathrm{t}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(0,n_\mathrm{t}-1). \end{aligned}$$

where the second equality follows from the definition of N, hence Equation (11).

In state $(n_\mathrm{a},0)$ where $n_\mathrm{a}\geq 1$, the next event must be a completion in the admission queue, and by the same argument as above, we have

$$u(n_\mathrm{a},0) = \frac{-C}{\mu_\mathrm{a}} + u(n_\mathrm{a}-1,1).$$

which is equivalent to

$$u(n_\mathrm{a},0) = \frac{-C}{\mu_\mathrm{t}+\mu_\mathrm{a}} + \frac{\mu_\mathrm{t}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(n_\mathrm{a},0) + \frac{\mu_\mathrm{a}}{\mu_\mathrm{t}+\mu_\mathrm{a}} u(n_\mathrm{a}-1,1)$$

as obtained by substituting $n_{\rm t}=0$ in Equation (10).

Finally, in state $(n_{\rm a},n_{\rm t})$ where $n_{\rm a}\geq 1, n_{\rm t}\geq 1$, if the next event is a completion in the admission queue, the customer leaving the admission queue will join the treatment queue if $n_{\rm t}<N$, otherwise they balk. Thus,

$$u(n_{\rm a},n_{\rm t})=\frac{-C}{\mu_{\rm t}+\mu_{\rm a}}+\frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a}-1,n_{\rm t}+\mathbb{1}_{\{n_{\rm t}<N\}})+\frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a},n_{\rm t}-1).$$

Proof of Proposition 1

We prove each property by induction on $n_{\rm a}$ and $n_{\rm t}$, starting with property (2) and then proving properties (1) and (3). In each proof, at the induction step, we assume $N\geq 2$, otherwise for $N=1$ the proof follows trivially from the base case.

Proof of Property (2)

Base case ($n_{\rm a}=0$): We first show that $u(0,N)<u(0,N-1)<\cdots<u(0,0)$ which we establish by induction on $n_{\rm t}$.

(1) When $n_{\rm t}=0$, we compare $u(0,0)$ and $u(0,1)$. By Lemma 1, with $u(0,0)=R-C/\mu_{\rm a}-C/\mu_{\rm t}>0$, we have

$$\begin{aligned}u(0,1)&=\frac{-C}{\mu_{\rm t}+\mu_{\rm a}}+\frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}\left[R-\frac{2}{\mu_{\rm t}}C\right]^{+}+\frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,0)\\&<\frac{-C}{\mu_{\rm t}+\mu_{\rm a}}+\frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}\left(R-\frac{C}{\mu_{\rm t}}\right)+\frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}\left(R-\frac{C}{\mu_{\rm t}}-\frac{C}{\mu_{\rm a}}\right)\\&=R-\frac{C}{\mu_{\rm a}}-\frac{C}{\mu_{\rm t}}=u(0,0).\end{aligned}$$

(2) Fix $n_{\rm t}\leq N-1$ and suppose that $u(0,n_{\rm t})<u(0,n_{\rm t}-1)$. We next show this implies $u(0,n_{\rm t}+1)<u(0,n_{\rm t})$. Recall by the definition of N that $R\geq NC/\mu_{\rm t}$, hence $R\geq (n_{\rm t}+1)C/\mu_{\rm t}$. Thus,

$$\begin{aligned}u(0,n_{\rm t}+1)&=-\frac{C}{\mu_{\rm t}+\mu_{\rm a}}+\frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t})+\frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}\left[R-\frac{n_{\rm t}+2}{\mu_{\rm t}}C\right]^{+}\\&\leq-\frac{C}{\mu_{\rm t}+\mu_{\rm a}}+\frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t})+\frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}\left(R-\frac{n_{\rm t}+1}{\mu_{\rm t}}C\right)\\&<-\frac{C}{\mu_{\rm t}+\mu_{\rm a}}+\frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}-1)+\frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}\left(R-\frac{n_{\rm t}+1}{\mu_{\rm t}}C\right)\\&=u(0,n_{\rm t}),\end{aligned}$$

where the inequality follows the hypothesis $u(0,n_{\rm t})<u(0,n_{\rm t}-1)$. Hence, $u(0,n_{\rm t})>u(0,n_{\rm t}+1)$ for every $n_{\rm t}\leq N-1$.

Induction step: Fix $n_{\rm a}\geq 0$ and assume the induction hypothesis that $u(n_{\rm a},N)<u(n_{\rm a},N-1)<\cdots<u(n_{\rm a},0)$. We shall show this implies $u(n_{\rm a}+1,N)<u(n_{\rm a}+1,N-1)<\cdots<u(n_{\rm a}+1,0)$. We prove this by induction on $n_{\rm t}$.

(1) When $n_{\mathrm{t}}=0$, we need to show that $u(n_{\mathrm{a}}+1,1)<u(n_{\mathrm{a}}+1,0)$. Noting that $u(n_{\mathrm{a}}+1,0)=-C/\mu_{\mathrm{a}}+u(n_{\mathrm{a}},1)$, from Lemma 1,

$$\begin{aligned} u(n_{\mathrm{a}}+1,1) &= \frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}},2)+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,0) \\ &< \frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}},1)+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}\left(\frac{-C}{\mu_{\mathrm{a}}}+u(n_{\mathrm{a}},1)\right) \\ &= \frac{-C}{\mu_{\mathrm{a}}}+u(n_{\mathrm{a}},1)=u(n_{\mathrm{a}}+1,0), \end{aligned}$$

where the inequality follows the hypothesis on n_{a}, i.e., $u(n_{\mathrm{a}},2)<u(n_{\mathrm{a}},1)$.

(2) Fix $n_{\mathrm{t}}\le N-1$ and suppose that $u(n_{\mathrm{a}}+1,n_{\mathrm{t}})<u(n_{\mathrm{a}}+1,n_{\mathrm{t}}-1)$. We next show this implies $u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+1)<u(n_{\mathrm{a}}+1,n_{\mathrm{t}})$. Indeed, by Lemma 1,

$$\begin{aligned} u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+1) &= \frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}},n_{\mathrm{t}}+1+\mathbb{1}_{\{n_{\mathrm{t}}+1<N\}})+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,n_{\mathrm{t}}) \\ &< \frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}},n_{\mathrm{t}}+1+\mathbb{1}_{\{n_{\mathrm{t}}+1<N\}})+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,n_{\mathrm{t}}-1) \\ &\le \frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}},n_{\mathrm{t}}+\mathbb{1}_{\{n_{\mathrm{t}}<N\}})+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,n_{\mathrm{t}}-1) \\ &= u(n_{\mathrm{a}}+1,n_{\mathrm{t}}). \end{aligned}$$

The first inequality above follows the hypothesis on n_{t}, i.e., $u(n_{\mathrm{a}}+1,n_{\mathrm{t}})<u(n_{\mathrm{a}}+1,n_{\mathrm{t}}-1)$. The second inequality follows because for $n_{\mathrm{t}}=N-1$,

$$u(n_{\mathrm{a}},n_{\mathrm{t}}+1+\mathbb{1}_{\{n_{\mathrm{t}}+1<N\}})=u(n_{\mathrm{a}},N)=u(n_{\mathrm{a}},n_{\mathrm{t}}+\mathbb{1}_{\{n_{\mathrm{t}}<N\}}),$$

whereas for $n_{\mathrm{t}}<N-1$, the hypothesis on n_{a}, that is $u(n_{\mathrm{a}},N)<\cdots<u(n_{\mathrm{a}},0)$, implies

$$u(n_{\mathrm{a}},n_{\mathrm{t}}+1+\mathbb{1}_{\{n_{\mathrm{t}}+1<N\}})=u(n_{\mathrm{a}},n_{\mathrm{t}}+2)<u(n_{\mathrm{a}},n_{\mathrm{t}}+1)=u(n_{\mathrm{a}},n_{\mathrm{t}}+\mathbb{1}_{\{n_{\mathrm{t}}<N\}}).$$

This concludes the proof of the second property in Proposition 1.

Proof of Property (1).

Base case ($n_{\mathrm{a}}=0$): We first show that $u(1,n_{\mathrm{t}})<u(0,n_{\mathrm{t}})$ for every $n_{\mathrm{t}}\le N$ which we establish induction on n_{t}.

(1) When $n_{\mathrm{t}}=0$, we need to show that $u(1,0)<u(0,0)$. By Lemma 1 we obtain

$$u(1,0)=\frac{-C}{\mu_{\mathrm{a}}}+u(0,1)<\frac{-C}{\mu_{\mathrm{a}}}+u(0,0)<u(0,0),$$

where the first inequality follows from part (2) which we proved previously.

(2) Fix $n_{\rm t} \le N-1$ and suppose that $u(1,n_{\rm t}) < u(0,n_{\rm t})$. We next show this implies $u(1,n_{\rm t}+1) < u(0,n_{\rm t}+1)$. Note first that for any $n_{\rm t} \le N$, when joining at state $(0,n_{\rm t})$, a customer must wait at least the time of $n_{\rm t}+1$ treatment services to receive the reward, which in expectation takes $(n_{\rm t}+1)/\mu_{\rm t}$ units of time, hence $u(0,n_{\rm t}) \le [R-(n_{\rm t}+1)C/\mu_{\rm t}]^+$. Using Lemma 1 and rearranging we then get

$$\begin{aligned} u(0,n_{\rm t}+1) &= \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}\left[R-\frac{n_{\rm t}+2}{\mu_{\rm t}}C\right]^+ + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}) \\ &= \frac{-C}{\mu_{\rm t}} + u(0,n_{\rm t}) + \frac{\mu_{\rm a}}{\mu_{\rm t}}\left(\left[R-\frac{n_{\rm t}+2}{\mu_{\rm t}}C\right]^+ - u(0,n_{\rm t}+1)\right) \\ &\ge \frac{-C}{\mu_{\rm t}} + u(0,n_{\rm t}). \end{aligned}$$

Thus, again by Lemma 1,

$$\begin{aligned} u(1,n_{\rm t}+1) &= \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}+1+\mathbb{1}_{\{n_{\rm t}+1<N\}}) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(1,n_{\rm t}) \\ &< \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}+1+\mathbb{1}_{\{n_{\rm t}+1<N\}}) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}) \\ &\le \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}+1) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}) \\ &= \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}+1) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}\left(\frac{-C}{\mu_{\rm t}} + u(0,n_{\rm t})\right) \\ &\le \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}+1) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(0,n_{\rm t}+1) = u(0,n_{\rm t}+1), \end{aligned}$$

where the first inequality follows the hypothesis on $n_{\rm t}$, namely $u(1,n_{\rm t}) < u(0,n_{\rm t})$, and the second inequality follows part (2) of the proposition which we proved previously.

<u>Induction step:</u> Fix $n_{\rm a} \ge 0$ and assume the induction hypothesis that $u(n_{\rm a}+1,n_{\rm t}) < u(n_{\rm a},n_{\rm t})$ for every $n_{\rm t} \le N$. We shall show this implies $u(n_{\rm a}+2,n_{\rm t}) < u(n_{\rm a}+1,n_{\rm t})$ for every $n_{\rm t} \le N$. We prove this by induction on $n_{\rm t}$.

(1) When $n_{\rm t}=0$, we need to show that $u(n_{\rm a}+2,0) < u(n_{\rm a}+1,0)$. By Lemma 1 we obtain

$$u(n_{\rm a}+2,0) = \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + u(n_{\rm a}+1,1) < \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + u(n_{\rm a}+1,0) < u(n_{\rm a}+1,0),$$

where the first inequality follows from part (2) of the proposition which we proved previously.

(2) Fix $n_{\rm t} \le N-1$ and suppose that $u(n_{\rm a}+2,n_{\rm t}) < u(n_{\rm a}+1,n_{\rm t})$. We next show this implies $u(n_{\rm a}+2,n_{\rm t}+1) < u(n_{\rm a}+1,n_{\rm t}+1)$. By Lemma 1,

$$\begin{aligned} u(n_{\rm a}+1,n_{\rm t}+1) &= \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a},n_{\rm t}+1+\mathbb{1}_{\{n_{\rm t}+1<N\}}) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a}+1,n_{\rm t}) \\ &> \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a},n_{\rm t}+1+\mathbb{1}_{\{n_{\rm t}+1<N\}}) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a}+2,n_{\rm t}) \\ &> \frac{-C}{\mu_{\rm t}+\mu_{\rm a}} + \frac{\mu_{\rm a}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a}+1,n_{\rm t}+1+\mathbb{1}_{\{n_{\rm t}+1<N\}}) + \frac{\mu_{\rm t}}{\mu_{\rm t}+\mu_{\rm a}}u(n_{\rm a}+2,n_{\rm t}) \\ &= u(n_{\rm a}+2,n_{\rm t}+1), \end{aligned}$$

where the first inequality follows the hypothesis on n_{t}, namely $u(n_{\mathrm{a}}+2,n_{\mathrm{t}})<u(n_{\mathrm{a}}+1,n_{\mathrm{t}})$, and the second inequality follows the hypothesis on n_{a}, namely, $u(n_{\mathrm{a}}+1,n_{\mathrm{t}})<u(n_{\mathrm{a}},n_{\mathrm{t}})$ for all $n_{\mathrm{t}}\le N$.

This concludes the proof of the first property in Proposition 1.

Proof of Property (3).

Base case ($n_{\mathrm{a}}=0$): We first show that $u(1,n_{\mathrm{t}})<u(0,n_{\mathrm{t}}+1)$ for every $n_{\mathrm{t}}\le N-1$ which we establish induction on n_{t}.

(1) When $n_{\mathrm{t}}=0$, we need to show that $u(1,0)<u(0,1)$. Indeed, by Lemma 1, $u(1,0)=-C/\mu_{\mathrm{a}}+u(0,1)<u(0,1)$.

(2) Fix $n_{\mathrm{t}}\le N-2$ and suppose that $u(1,n_{\mathrm{t}})<u(0,n_{\mathrm{t}}+1)$. We next show this implies $u(1,n_{\mathrm{t}}+1)<u(0,n_{\mathrm{t}}+2)$. By Lemma 1,

$$\begin{aligned}
u(1,n_{\mathrm{t}}+1)&=\frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(1,n_{\mathrm{t}})+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(0,n_{\mathrm{t}}+2)\\
&<\frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(0,n_{\mathrm{t}}+1)+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(0,n_{\mathrm{t}}+2)\\
&\le\frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(0,n_{\mathrm{t}}+1)+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}\left[R-\frac{n_{\mathrm{t}}+3}{\mu_{\mathrm{t}}}C\right]^{+}\\
&=u(0,n_{\mathrm{t}}+2),
\end{aligned}$$

where the first inequality follows the hypothesis on n_{t}, i.e., $u(1,n_{\mathrm{t}})<u(0,n_{\mathrm{t}}+1)$, and the second inequality follows from $u(0,n_{\mathrm{t}})\le[R-(n_{\mathrm{t}}+1)C/\mu_{\mathrm{t}}]^{+}$ for all $n_{\mathrm{t}}\le N$ as shown in the proof of part (1).

Induction step: Fix $n_{\mathrm{a}}\ge 0$ and assume the induction hypothesis that $u(n_{\mathrm{a}}+1,n_{\mathrm{t}})<u(n_{\mathrm{a}},n_{\mathrm{t}}+1)$ for every $n_{\mathrm{t}}\le N-1$. We shall show this implies $u(n_{\mathrm{a}}+2,n_{\mathrm{t}})<u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+1)$ for every $n_{\mathrm{t}}\le N$. We prove this by induction on n_{t}.

(1) When $n_{\mathrm{t}}=0$, we need to show that $u(n_{\mathrm{a}}+1,0)<u(n_{\mathrm{a}},1)$. Indeed, by Lemma 1, $u(n_{\mathrm{a}}+1,0)=-C/\mu_{\mathrm{a}}+u(n_{\mathrm{a}},1)<u(n_{\mathrm{a}},1)$.

(2) Fix $n_{\mathrm{t}}\le N-2$ and assume $u(n_{\mathrm{a}}+2,n_{\mathrm{t}})<u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+1)$. We shall show this implies $u(n_{\mathrm{a}}+2,n_{\mathrm{t}}+1)<u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+2)$

$$\begin{aligned}
u(n_{\mathrm{a}}+2,n_{\mathrm{t}}+1)&=\frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+2,n_{\mathrm{t}})+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+2)\\
&<\frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+1)+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+2)\\
&<\frac{-C}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}+\frac{\mu_{\mathrm{t}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+1)+\frac{\mu_{\mathrm{a}}}{\mu_{\mathrm{t}}+\mu_{\mathrm{a}}}u(n_{\mathrm{a}},n_{\mathrm{t}}+2+\mathbb{1}_{\{n_{\mathrm{t}}+2<N\}})\\
&=u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+2).
\end{aligned}$$

The first inequality follows the hypothesis on n_{t}, namely $u(n_{\mathrm{a}}+2,n_{\mathrm{t}})<u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+1)$. The second inequality follows because when $n_{\mathrm{t}}=N-2$, according to part (1) which we proved previously,

$$u(n_{\mathrm{a}}+1,n_{\mathrm{t}}+2)=u(n_{\mathrm{a}}+1,N)<u(n_{\mathrm{a}},N)=u(n_{\mathrm{a}},n_{\mathrm{t}}+2+\mathbb{1}_{\{n_{\mathrm{t}}+2<N\}}),$$

whereas for $n_\mathrm{t} < N-2$, the induction hypothesis on n_a implies

$$u(n_\mathrm{a}+1, n_\mathrm{t}+2) < u(n_\mathrm{a}, n_\mathrm{t}+3) = u(n_\mathrm{a}, n_\mathrm{t}+2+\mathbb{1}_{\{n_\mathrm{t}+2<N\}}).$$

This concludes the proof of the third property in Proposition 1.

Proof of Lemma 2

Recall that a customer, upon transitioning to the treatment queue, balks when they observe N customers in it. We first argue that when N customers (or more) are awaiting in treatment, a customer will not join the admission queue either.

Consider a tagged customer commencing service in the admission queue when there are N customers in treatment. Let this tagged customer's (remaining) service time in admission be S, and denote the (remaining) service times of the customers in treatment by $T_1, \ldots T_N$, with T_{N+1} being the service time of the tagged customer should they join the treatment queue. Then, assuming the customer in admission never balks, their wait time is given by:

$$S + \left[\sum_{i=1}^{N} T_i - S\right]^+ + T_{N+1} \geq S + \sum_{i=1}^{N} T_i - S + T_{N+1} = \sum_{i=1}^{N+1} T_i,$$

where the inequality holds with probability 1. Taking expectation on both sides and multiplying by C we see that the tagged customer's total waiting cost is bounded from below by $C(N+1)/\mu_\mathrm{t}$, where we have, by definition of N, that $C(N+1)/\mu_\mathrm{t} > R$. Hence, in equilibrium, a customer would not join the admission queue when the system is at state $(0, N)$. Furthermore, by diagonal monotonicty (Proposition 1 Item 3), joining at any state $(k, N-k)$, $k = 1, \ldots, N$ yields even less utility and therefore is undesired. It follows that in equilibrium, the total number of customers never exceeds N, and as a result, customers never balk when transitioning from admission to treatment.

Proof of Theorem 1

For each possible state $(n_\mathrm{a}, n_\mathrm{t})$, the expected utility $u(n_\mathrm{a}, n_\mathrm{t})$ is determined uniquely by Lemma 1. Therefore, a unique equilibrium (dominant) strategy exists, prescribing joining at state $(n_\mathrm{a}, n_\mathrm{t})$ if and only if $u(n_\mathrm{a}, n_\mathrm{t}) > 0$. In particular, if $(n_\mathrm{a}, n_\mathrm{t})$ satisfies $u(n_\mathrm{a}, n_\mathrm{t}) > 0$, then by Proposition 1 properties 1 and 2, $u(n_\mathrm{a}-k, n_\mathrm{t}-l) > 0$ for all $k \in [0, n_\mathrm{a}]$, $l \in [0, n_\mathrm{t}]$, hence, the equilibrium strategy is a two-dimensional threshold strategy.

Next, we prove the sum-strategy condition. By Proposition 1 property 3 (diagonal monotonicity) and the definition of $N_\mathrm{t}(0)$, for any $k \in [0, N_\mathrm{t}(0)]$

$$u(N_\mathrm{t}(0)-k, k) \leq u(0, N_\mathrm{t}(0)) \leq 0,$$

hence, customers do not join at state $u(N_\mathrm{t}(0)-k, k)$. Suppose that $u(N_\mathrm{t}(0)-1, 0) > 0$. Then once again using Proposition 1 property 3 we obtain $u(N_\mathrm{t}(0)-1-k, k) \geq u(N_\mathrm{t}(0)-1, 0) > 0$ for all $k \in [0, N_\mathrm{t}(0)-1]$. Thus, in equilibrium, a customer joins the system at state $(n_\mathrm{a}, n_\mathrm{t})$ if and only $n_\mathrm{a} + n_\mathrm{t} \leq N_\mathrm{t}(0) - 1$.

Proof of Proposition 2

The fact that customers do not renege from the treatment queue is directly implied by Naor (1969). In the proof we shall show that a customer will not renege from the admission queue as well.

Consider a customer who is waiting in the admission queue at time t_0, and let $t_1 \geq t_0$ be a time instant when the customer was still in the admission queue. From this customer's perspective, the length of the treatment queue at times t_0 and t_1 are a random variable, which we denote by Y^{t_0} and Y^{t_1}, respectively. Let $n_{\rm a}^{t_0}$ and $n_{\rm a}^{t_1}$ denote the number of other customers ahead of the tagged customer in the admission queue at time instants t_0 and t_1, respectively. We recall that by assumption, the tagged customer observes both $n_{\rm a}^{t_0}$ and $n_{\rm a}^{t_1}$. Consider an arbitrary time point $t_1 > t_0$ such that the customer was still in the admission queue. To prove that this customer will not renege, we need to show that

$$\mathbb{E}\left[u(n_{\rm a}^{t_0}, Y^{t_0})\right] \leq \mathbb{E}\left[u(n_{\rm a}^{t_1}, Y^{t_1}),\right], \tag{12}$$

where the expectations in the left- and right-hand sides are taken over the random variables Y^{t_1} and Y^{t_0}, respectively.

Among those who arrived before the discussed customer, let $d = n_{\rm a}^{t_0} - n_{\rm a}^{t_1} \geq 0$ represent the number of customers who departed from the admission queue during time interval $(t_0, t_1]$. Out of those customers, let B represent the number of customers who left the admission queue but did not join the treatment queue (that is, either balked or reneged). Thus, B is a random variable taking values in $\{0, 1, \cdots, d\}$. Let D be the random variable which represents the number of departures from the treatment queue during the time interval $(t_0, t_1]$. Hence, we have $Y^{t_1} = Y^{t_0} + d - B - D \geq 0$ (with probability 1). By nonnegativity of B and D, we have

$$\left(Y^{t_0} - B - D\right)^+ \leq_{\rm st} Y^{t_0},$$

Therefore, because $u(n_{\rm a}, n_{\rm t})$ is non-increasing in $n_{\rm t}$ as shown in Proposition 1, we have

$$\mathbb{E}\left[u\left(n_{\rm a}^{t_0}, Y^{t_0}\right)\right] \leq \mathbb{E}\left[u\left(n_{\rm a}^{t_0}, \left(Y^{t_0} - B - D\right)^+\right)\right]. \tag{13}$$

In what follows, we claim that

$$\mathbb{E}\left[u\left(n_{\rm a}^{t_0}, \left(Y^{t_0} - B - D\right)^+\right)\right] \leq \mathbb{E}\left[u(n_{\rm a}^{t_1}, Y^{t_1})\right]. \tag{14}$$

Together with Eq. (13), we then establish Eq. (12). We shall now prove Eq. (14). From the Law of Total Probability we have

$$\begin{aligned}\mathbb{E}\left[u\left(n_{\rm a}^{t_0}, \left(Y^{t_0} - B - D\right)^+\right)\right] &= \mathbb{E}\left[u\left(n_{\rm a}^{t_0}, \left(Y^{t_1} - d\right)^+\right)\right] \\ &= \sum_{0 \leq m \leq d} u(n_{\rm a}^{t_0}, 0)\mathbb{P}[Y^{t_1} = m] + \sum_{m > d} u(n_{\rm a}^{t_0}, m - d)\mathbb{P}[Y^{t_1} = m],\end{aligned}$$

and

$$\mathbb{E}\left[u(n_{\mathrm{a}}^{t_1}, Y^{t_1})\right] = \mathbb{E}\left[u(n_{\mathrm{a}}^{t_0} - d, Y^{t_1})\right] = \sum_{m\geq 0} u(n_{\mathrm{a}}^{t_0} - d, m)\mathbb{P}[Y^{t_1} = m].$$

To establish Eq. (14), we show that $u(n_{\mathrm{a}}^{t_0} - d, m) \geq u(n_{\mathrm{a}}^{t_0}, 0)$ if $m \leq d$ and $u(n_{\mathrm{a}}^{t_0} - d, m) \geq u(n_{\mathrm{a}}^{t_0}, m-d)$ if $m > d$:

Case I: If $m \leq d$, by diagonal monotonicity of $u(n_{\mathrm{a}}, n_{\mathrm{t}})$ shown in Proposition 1, we have $u(n_{\mathrm{a}}^{t_0} - d, m) \geq u(n_{\mathrm{a}}^{t_0} - d + 1, m-1) \geq \cdots \geq u(n_{\mathrm{a}}^{t_0} - d + m, 0)$. If $m = d$, then we are done; otherwise if $m < d$, we further have $u(n_{\mathrm{a}}^{t_0} - d + m, 0) \geq u(n_{\mathrm{a}}^{t_0} - d + m + 1, 0) \geq \cdots \geq u(n_{\mathrm{a}}^{t_0}, 0)$, because $u(n_{\mathrm{a}}, n_{\mathrm{t}})$ is non-increasing in n_{a} as shown in Proposition 1.

Case II: If $m > d$, again by diagonal monotonicity, we have $u(n_{\mathrm{a}}^{t_0} - d, m) \geq u(n_{\mathrm{a}}^{t_0} - d + 1, m-1) \geq \cdots \geq u(n_{\mathrm{a}}^{t_0}, m-d)$.

Proof of Proposition 3

The proof of Proposition 3 relies on Lemma 5 which we state and prove below. Throughout the proof, we assume $\boldsymbol{p}$ as arbitrarily fixed, therefore we omit it from the notation. Unless otherwise specified, when comparing random variables defined on the same probability space using standard equality/inequality signs, we mean 'with probability 1'. For a random variable Z and an event A we consider $[Z \mid A]$ an arbitrary random variable whose distribution is the conditional distribution of Z given A.

To state Lemma 5, let $\{(X(t), Y(t))\}_{t\geq 0}$ represent the system state process over time, with $X(t)$ and $Y(t)$ being the admission and treatment queue lengths at time t, respectively. To clarify, we make the typical assumption that $X(t)$ and $Y(t)$ are right-continuous with left limits (RCLL), thus, if t is a time of an event in the system, $(X(t), Y(t))$ is the system state after that event materializes. We denote the left limits of $X(\cdot)$ and $Y(\cdot)$ at t by $X(t-)$ and $Y(t-)$, respectively. Let also the pair of random variables (X, Y) represent the stationary counterpart of $(X(t), Y(t))$, meaning that $(X(t), Y(t))$ converges to (X, Y) in law as $t \to \infty$.

LEMMA 5. *For every state* $n_{\mathrm{a}} \in \{0, ..., N_{\mathrm{a}}(\boldsymbol{p}) - 1\}$,

$$[Y \mid X = n_{\mathrm{a}}] \leq_{\mathrm{st}} [Y \mid X = n_{\mathrm{a}} + 1] + 1, \tag{15}$$

where $\leq_{\mathrm{st}}$ *denotes inequality in the usual stochastic order sense.*

Proof: We prove Lemma 5 by induction on n_{a}, starting with the base case, $n_{\mathrm{a}} = 0$. Suppose the system is stationary and assume at some arbitrary instant t, a customer arrives and finds 0 other customers in admission (so that $X(t-) = 0$). Note that an arrival to the system at time t does not change the length at the treatment queue, thus, $Y(t) = Y(t-) =_{\mathrm{st}} [Y \mid X = 0]$, where $=_{\mathrm{st}}$ denotes equality in law. Let $t' \in (-\infty, t)$ denote the time of the last admission-service completion prior to time t. Thus, a moment before t', the length of the admission queue was 1, namely, $X(t'-) = 1$. Because the system process is stationary and

Markov, it can be assumed that at that moment, shortly before t', the (random) treatment queue length was stationary, i.e., $Y(t'-) =_{\text{st}} [Y \mid X = 1]$. Note that at time t', the treatment-queue length increases by 1 if (and only if) $Y(t'-) < N$. Denote by $D_{(t',t)}$ the number of departures from the treatment queue during the time interval (t', t) which is non-negative. Hence, we have

$$[Y \mid X = 0] =_{\text{st}} Y(t) = Y(t'-) + \mathbb{1}_{\{Y(t'-)<N\}} - D_{(t',t)} < Y(t'-) + 1 =_{\text{st}} [Y \mid X = 1] + 1,$$

which proves the base case.

For the induction step, consider now $n_{\text{a}} \in \{1, \ldots, N_{\text{a}}(\boldsymbol{p}) - 1\}$, and assume the induction hypothesis that

$$[Y \mid X = n_{\text{a}} - 1] \leq_{\text{st}} [Y \mid X = n_{\text{a}}] + 1.$$

We next show this implies Equation (15). As before, suppose a customer arrives at the stationary system at time t and finds n_{a} customers in admission, thus, $X(t-) = n_{\text{a}}$ and $Y(t) =_{\text{st}} [Y \mid X = n_{\text{a}}]$. Let now $t' \in (-\infty, t)$ denote the last time prior to t that the admission queue transitioned into level n_{a} from a different state, so that during $[t', t)$ this level remained constant at n_{a}. Thus, $X(t'-)$ is a random variable taking one out of two possible values, either $n_{\text{a}} - 1$ or $n_{\text{a}} + 1$. The event $X(t'-) = n_{\text{a}} - 1$ corresponds to the joining at time t of a customer who observes an admission queue length of $n_{\text{a}} - 1$, whereas $X(t'-) = n_{\text{a}} + 1$ corresponds to an admission-service completion at time t assuming a moment right before it the admission queue length was $n_{\text{a}} + 1$. Note that at time t', in the former case ($X(t'-) = n_{\text{a}} - 1$), the treatment queue length remains unchanged, while in the latter case ($X(t'-) = n_{\text{a}} + 1$), one customer joins the treatment queue if (and only if) its length is less than N. Denoting by $D_{(t',t)}$ the number of departures from the treatment queue during (t', t), we have

$$\begin{aligned}
Y(t) &= \mathbb{1}_{\{X(t'-)=n_{\text{a}}-1\}} Y(t'-) + \mathbb{1}_{\{X(t'-)=n_{\text{a}}+1\}} \left(Y(t'-) + \mathbb{1}_{\{Y(t'-)<N\}} \right) - D_{(t',t)} \\
&\leq \mathbb{1}_{\{X(t'-)=n_{\text{a}}-1\}} Y(t'-) + \mathbb{1}_{\{X(t'-)=n_{\text{a}}+1\}} \left(Y(t'-) + 1 \right) \\
&=_{\text{st}} Z \cdot [Y \mid X = n_{\text{a}} - 1] + (1 - Z) \left([Y \mid X = n_{\text{a}} + 1] + 1 \right),
\end{aligned}$$

where Z is an independent Bernoulli random variable distributed as $\mathbb{1}_{\{X(t'-)=n_{\text{a}}-1\}}$. This, together with $Y(t) =_{\text{st}} [Y \mid X = n_{\text{a}}]$ and the induction hypothesis implies

$$[Y \mid X = n_{\text{a}}] \leq_{\text{st}} Z \cdot \left([Y \mid X = n_{\text{a}}] + 1 \right) + (1 - Z) \left([Y \mid X = n_{\text{a}} + 1] + 1 \right).$$

Clearly, $[Y \mid X = n_{\text{a}}] =_{\text{st}} Z \cdot [Y \mid X = n_{\text{a}}] + (1 - Z) \cdot [Y \mid X = n_{\text{a}}]$, so that the latter inequality is equivalent to

$$Z \cdot [Y \mid X = n_{\text{a}}] + (1 - Z) \cdot [Y \mid X = n_{\text{a}}] \leq_{\text{st}} Z \cdot [Y \mid X = n_{\text{a}}] + (1 - Z) \left([Y \mid X = n_{\text{a}} + 1] + 1 \right) + Z.$$

Thus, we can assume without loss of generality that Z, $[Y \mid X = n_{\text{a}}]$ and $[Y \mid X = n_{\text{a}} + 1]$ are all defined on the same probability space (see, e.g., Theorem 1.A.1 in Shaked and Shanthikumar (2007)), such that

$$(1 - Z) \cdot [Y \mid X = n_{\text{a}}] \leq (1 - Z) \left([Y \mid X = n_{\text{a}} + 1] + 1 \right) + Z.$$

with probability 1. Conditioning on $\{Z=0\}$ implies that there exists a probability space in which (with probability 1)

$$[Y \mid X = n_{\rm a}] \le [Y \mid X = n_{\rm a}+1]+1,$$

therefore proving Lemma 5. □

Based on Lemma 5, we move on to proving Proposition 3. By Lemma 5, there exists a probability space in which, with probability 1, $[Y \mid X = n_{\rm a}] \le [Y \mid X = n_{\rm a}+1]+1$, and to show our claim, we can restrict attention to this probability space without loss of generality. Because $u(\cdot,\cdot)$ is non-increasing in its second argument (Proposition 1), we therefore have (with probability 1),

$$u(n_{\rm a}, [Y \mid X = n_{\rm a}+1]+1) \le u(n_{\rm a}, [Y \mid X = n_{\rm a}]),$$

and by the diagonal monotonicity shown in Proposition 1, we further have

$$u(n_{\rm a}+1, [Y \mid X = n_{\rm a}+1]) < u(n_{\rm a}, [Y \mid X = n_{\rm a}+1]+1).$$

Combining these two inequalities, we see that

$$u(n_{\rm a}+1, [Y \mid X = n_{\rm a}+1]) < u(n_{\rm a}, [Y \mid X = n_{\rm a}]).$$

Taking expectation on both sides (with respect to the joint measure of $[Y \mid X = n_{\rm a}]$ and $[Y \mid X = n_{\rm a}+1]$ in the aforementioned probability space) yields $u_{\boldsymbol{p}}(n_{\rm a}+1) < u_{\boldsymbol{p}}(n_{\rm a})$.

Proof of Theorem 2

Part I - Existence

The crux of the existence proof is to apply Kakutani's fixed-point theorem (see, e.g. Smart (1980)) on the best-response correspondence, which we will define below. This requires proving that this best-response correspondence is upper-hemicontinuous. Recall that a correspondence (i.e., a set-valued function) $\mathcal{G}: X \to 2^{\mathcal{Y}}$ is upper-hemicontinuous at $x \in \mathcal{X}$ if for any open set $\mathcal{R} \subset \mathcal{Y}$ with $\mathcal{G}(x) \subset \mathcal{R}$, there exists a neighborhood $\mathcal{U}$ of x, such that for all $x' \in \mathcal{U}$, $\mathcal{G}(x') \subset \mathcal{R}$. We say that $\mathcal{G}$ is upper-hemicontinuous on $\mathcal{X}$ if $\mathcal{G}$ is upper-hemicontinuous at each $x \in \mathcal{X}$.

Denote by $\mathcal{S} = [0,1]^{N_{\rm a}(0)}$ the strategy space, which is compact and convex. Define the best-response correspondence $\mathcal{BR}: \mathcal{S} \to 2^{\mathcal{S}}$ as

$$\mathcal{BR}(\boldsymbol{p}) = \left\{ \boldsymbol{q} \in \mathcal{S} \text{ s.t. } q_{n_{\rm a}} = \begin{cases} 1 \text{ if } u_{\boldsymbol{p}}(n_{\rm a}) > 1, \\ 0 \text{ if } u_{\boldsymbol{p}}(n_{\rm a}) < 0, \end{cases} \quad n_{\rm a} = 0, \ldots, N_{\rm a}(0)-1 \right\} \tag{16}$$

By definition, any fixed point $\boldsymbol{p} \in \mathcal{BR}(\boldsymbol{p})$ of the best-response correspondence is an equilibrium.

We start with restating Katutani's fixed-point theorem.

THEOREM 3 (**Kakutani's Fixed Point Theorem**). *Let $\mathcal{X} \subset \mathbb{R}^n$ be compact and convex. If $\mathcal{G} : \mathcal{X} \to 2^{\mathcal{X}}$ is an upper-hemicontinuous correspondence that has nonempty, compact, and convex values, then $\mathcal{G}$ has a fixed point.*

From Eq. (16) it is clear that $\mathcal{BR}(\boldsymbol{p})$ is compact, convex and non empty for all $\boldsymbol{p} \in \mathcal{S}$. It suffices therefore to show that $\mathcal{BR}$ is upper-hemicontinuous. Let $\bar{u}(\boldsymbol{q}, \boldsymbol{p}) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be the ex-ante expected utility for a customer who adopts strategy $\boldsymbol{q}$ when all other customers follow $\boldsymbol{p}$. Thus,

$$\bar{u}(\boldsymbol{q}, \boldsymbol{p}) = \sum_{n_{\mathrm{a}}=0}^{N_{\mathrm{a}}(\boldsymbol{p})} \pi_{\boldsymbol{p}}(n_{\mathrm{a}}, \cdot) \cdot u_{\boldsymbol{p}}(n_{\mathrm{a}}) \cdot q_{n_{\mathrm{a}}}.$$

The best-response correspondence can then be equivalently defined as $\mathcal{BR}(\boldsymbol{p}) = \arg\max_{\boldsymbol{q} \in \mathcal{S}} \; \bar{u}(\boldsymbol{q}, \boldsymbol{p})$. From the Maximum Theorem, if $\bar{u}(\boldsymbol{q}, \boldsymbol{p})$ is continuous in $(\boldsymbol{q}, \boldsymbol{p})$, then $\mathcal{BR}$ is compact-valued and upper-hemicontinuous, thereby proving the existence of the equilibrium strategy, in accordance with Theorem 3. The rest of the proof is therefore dedicated to proving the continuity of $\bar{u}$ that we state as Lemma 8 below. We remark one subtle yet significant challenge in the proof. For a given strategy $\boldsymbol{p}$, it is possible that a very small perturbation can cause a dramatic change of the induced recurrent states. For instance, perturbing $\boldsymbol{p}$ from $[1, \delta, 1, 1, 0]$ to $[1, 0, 1, 1, 0]$ (δ close to zero) will change the states $\{2, 3, 4\}$ from recurrent to transient states. Therefore, the expected utility will be computed over a possibly very different landscape of states. To overcome this issue, we must carefully and precisely characterize the underlying continuity.

Towards proving Lemma 8 we make several observations. Let $\mathcal{N} = \{0, \ldots, N_{\mathrm{a}}(0)\} \times \{0, \ldots, N\}$ be the state space of the underlying Markov process, and let Π be the mapping of a strategy $\boldsymbol{p} \in \mathcal{S}$ to the (unique) set of stationary probabilities $\{\pi_{\boldsymbol{p}}(n_{\mathrm{a}}, n_{\mathrm{t}})\}_{(n_{\mathrm{a}}, n_{\mathrm{t}}) \in \mathcal{N}}$ solving the following balance equations:

$$\begin{aligned}\left(\mu_{\mathrm{t}} \mathbb{1}_{\{n_{\mathrm{t}}>0\}} + \mu_{\mathrm{a}} \mathbb{1}_{\{n_{\mathrm{a}}>0\}} + \lambda p_{n_{\mathrm{a}}} \mathbb{1}_{\{n_{\mathrm{a}}<N_{\mathrm{a}}(\boldsymbol{p})\}}\right) \pi_{\boldsymbol{p}}(n_{\mathrm{a}}, n_{\mathrm{t}}) = {} & \lambda p_{n_{\mathrm{a}}-1} \pi_{\boldsymbol{p}}(n_{\mathrm{a}}-1, n_{\mathrm{t}}) \mathbb{1}_{\{n_{\mathrm{a}}>0\}} \\ & + \mu_{\mathrm{a}} \pi_{\boldsymbol{p}}(n_{\mathrm{a}}+1, n_{\mathrm{t}}-1) \mathbb{1}_{\{n_{\mathrm{a}}<N_{\mathrm{a}}(\boldsymbol{p}), n_{\mathrm{t}}>0\}} \\ & + \mu_{\mathrm{a}} \pi_{\boldsymbol{p}}(n_{\mathrm{a}}+1, N) \mathbb{1}_{\{n_{\mathrm{a}}<N_{\mathrm{a}}(\boldsymbol{p}), n_{\mathrm{t}}=N\}} \\ & + \mu_{\mathrm{t}} \pi_{\boldsymbol{p}}(n_{\mathrm{a}}, n_{\mathrm{t}}+1) \mathbb{1}_{\{n_{\mathrm{t}}<N\}}.\end{aligned} \tag{17}$$

LEMMA 6. *The mapping Π is continuous in $\boldsymbol{p}$.*

Proof: We first restate a result on parametric continuity of stationary probabilities of discrete-time Markov chains (DTMC) from Le Van and Stachurski (2007).

THEOREM 4 (**Theorem 4.1 in Le Van and Stachurski (2007)**). *Let θ be a vector in $\mathbb{R}^d$. Consider a finite-state DTMC with transition probability matrix $P(\theta)$, where each element is a continuous function of θ. If the stationary probability μ uniquely exists, then the mapping $\theta \to \mu(\theta)$ is continuous on $\mathbb{R}^d$.*

Consider some fixed strategy $\boldsymbol{p} \in \mathcal{S}$ and let $Q \in \mathbb{R}^{(N_{\rm a}(0)+1)\times(N+1)}$ be the transition-rate matrix of the original CTMC when customers join according to $\boldsymbol{p}$, thus, its unique stationary distribution is $\Pi(\boldsymbol{p})$. From the balance equations in (17), it can be seen that each entry of Q is a linear function of $\boldsymbol{p}$ and therefore is continuous in $\boldsymbol{p}$. Furthermore, each such transition rate is at most $\gamma = \lambda + \mu_{\rm a} + \mu_{\rm t}$. Define the uniformized DTMC by constructing the one-step transition probability matrix $P = I + \gamma^{-1}Q$, where I is the identity matrix. Thus, each entry of P is continuous in $\boldsymbol{p}$ (see, e.g., Grassmann (1977)). As Q is ergodic, the uniformized DTMC characterized by P admits a unique stationary distribution which is identical to $\Pi(\boldsymbol{p})$. By Theorem 4, this implies that $\Pi(\boldsymbol{p})$, hence proving Lemma 6. $\square$

The next lemma shows that for any strategy $\boldsymbol{p}$ and any admission-queue length $n_{\rm a}$ that customers may observe under $\boldsymbol{p}$, the expected utility $u_{\boldsymbol{p}}(n_{\rm a})$ is continuous at $\boldsymbol{p}$. In other words, a small deviation from $\boldsymbol{p}$ to another strategy results in a small change in $u_{\boldsymbol{p}}(n_{\rm a})$, so long as it does not change the support of $\Pi(\boldsymbol{p})$. For the remainder of the proof, it is useful to note that, by Proposition 1, for any state $(n_{\rm a}, n_{\rm t}) \in \mathcal{S}$, $u(N_{\rm a}(0), N) \le u(n_{\rm a}, n_{\rm t}) \le u(0,0)$, hence with $U = \max\{|u(N_{\rm a}(0), N)|, u(0,0)\}$ we have $|u(n_{\rm a}, n_{\rm t})| \le U$ for all $(n_{\rm a}, n_{\rm t}) \in \mathcal{S}$.

LEMMA 7. *Consider a strategy $\boldsymbol{p} \in \mathcal{S}$. For any $\epsilon > 0$, there exists $\delta > 0$, such that for any strategy $\boldsymbol{p}'$ satisfying $\|\boldsymbol{p} - \boldsymbol{p}'\| < \delta$ and any state $n_{\rm a} \le \min\{N_{\rm a}(\boldsymbol{p}), N_{\rm a}(\boldsymbol{p}')\}$, we have $|\, u_{\boldsymbol{p}}(n_{\rm a}) - u_{\boldsymbol{p}'}(n_{\rm a})\,| < \epsilon$.*

Proof: Set $\epsilon > 0$ and denote $\pi_{\min} = \min_{n_{\rm a} \le N_{\rm a}(\boldsymbol{p})} \pi_{\boldsymbol{p}}(n_{\rm a}, \cdot)$, hence, $\pi_{\min} > 0$. By the continuity of Π shown in Lemma 6, we can choose δ sufficiently small such that for any $\|\boldsymbol{p} - \boldsymbol{p}'\| < \delta$,

$$\|\Pi(\boldsymbol{p}) - \Pi(\boldsymbol{p}')\| < \frac{\pi_{\min}^2}{U \cdot (N+2)^2}\epsilon.$$

Then, for any $n_{\rm a} < \min\{N_{\rm a}(\boldsymbol{p}), N_{\rm a}(\boldsymbol{p}')\}$ we have

$$\begin{aligned}
|\, u_{\boldsymbol{p}}(n_{\rm a}) - u_{\boldsymbol{p}'}(n_{\rm a})\,| &= \left| \sum_{n_{\rm t}=0}^{N} \left(\frac{\pi_{\boldsymbol{p}}(n_{\rm a}, n_{\rm t})}{\pi_{\boldsymbol{p}}(n_{\rm a}, \cdot)} - \frac{\pi_{\boldsymbol{p}'}(n_{\rm a}, n_{\rm t})}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)} \right) u(n_{\rm a}, n_{\rm t}) \right| \\
&= \left| \sum_{n_{\rm t}=0}^{N} \left(\left(\frac{1}{\pi_{\boldsymbol{p}}(n_{\rm a}, \cdot)} - \frac{1}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)} \right) \pi_{\boldsymbol{p}}(n_{\rm a}, n_{\rm t}) + \frac{\pi_{\boldsymbol{p}}(n_{\rm a}, n_{\rm t}) - \pi_{\boldsymbol{p}'}(n_{\rm a}, n_{\rm t})}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)} \right) u(n_{\rm a}, n_{\rm t}) \right| \\
&\le \sum_{n_{\rm t}=0}^{N} \left(\left| \frac{1}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)} - \frac{1}{\pi_{\boldsymbol{p}}(n_{\rm a}, \cdot)} \right| + \frac{|\pi_{\boldsymbol{p}}(n_{\rm a}, n_{\rm t}) - \pi_{\boldsymbol{p}'}(n_{\rm a}, n_{\rm t})|}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)} \right) |u(n_{\rm a}, n_{\rm t})| \\
&\le U \cdot (N+1) \cdot \left(\left| \frac{1}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)} - \frac{1}{\pi_{\boldsymbol{p}}(n_{\rm a}, \cdot)} \right| + \frac{\|\Pi(\boldsymbol{p}) - \Pi(\boldsymbol{p}')\|}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)} \right) \\
&\le U \cdot (N+1) \cdot \left(\left| \frac{\pi_{\boldsymbol{p}}(n_{\rm a}, \cdot) - \pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)\pi_{\boldsymbol{p}}(n_{\rm a}, \cdot)} \right| + \frac{\|\Pi(\boldsymbol{p}) - \Pi(\boldsymbol{p}')\|}{\pi_{\boldsymbol{p}'}(n_{\rm a}, \cdot)^2} \right) \\
&\le U \cdot (N+1) \cdot \left(\frac{\sum_{n_{\rm t}=0}^{N} |\pi_{\boldsymbol{p}}(n_{\rm a}, n_{\rm t}) - \pi_{\boldsymbol{p}'}(n_{\rm a}, n_{\rm t})|}{\pi_{\min}^2} + \frac{\|\Pi(\boldsymbol{p}) - \Pi(\boldsymbol{p}')\|}{\pi_{\min}^2} \right) \\
&\le U \cdot (N+1) \cdot \left(\frac{(N+1)\|\Pi(\boldsymbol{p}) - \Pi(\boldsymbol{p}')\|}{\pi_{\min}^2} + \frac{\|\Pi(\boldsymbol{p}) - \Pi(\boldsymbol{p}')\|}{\pi_{\min}^2} \right) \\
&\le \frac{U \cdot (N+2)^2}{\pi_{\min}^2} \|\Pi(\boldsymbol{p}) - \Pi(\boldsymbol{p}')\| < \epsilon
\end{aligned}$$

□

LEMMA 8. *The ex-ante expected utility $\bar{u}(\boldsymbol{q},\boldsymbol{p})$ is continuous in $(\boldsymbol{q},\boldsymbol{p})$.*

Proof: Set $\epsilon > 0$ and $(\boldsymbol{q},\boldsymbol{p})$, then by Lemmas 6 and 7, we can choose some sufficiently small δ such that for all $(\boldsymbol{q}',\boldsymbol{p}')$ satisfying $\|(\boldsymbol{q},\boldsymbol{p})-(\boldsymbol{q}',\boldsymbol{p}')\|<\delta$, the following inequalities hold:

$$\|\boldsymbol{p}-\boldsymbol{p}'\|\leq\frac{\epsilon}{4UN},\quad |u_{\boldsymbol{p}}(n_{\rm a})-u_{\boldsymbol{p}'}(n_{\rm a})|\leq\frac{\epsilon}{4N},\quad \|\Pi(\boldsymbol{p})-\Pi(\boldsymbol{p}')\|\leq\frac{\epsilon}{4U(N+1)N},\quad \|\boldsymbol{q}-\boldsymbol{q}'\|\leq\frac{\epsilon}{4UN}.$$

Consider such δ and $(\boldsymbol{q}',\boldsymbol{p}')$ as above and assume without loss of generality that $N_{\rm a}(\boldsymbol{p})\leq N_{\rm a}(\boldsymbol{p}')$.

To prove that $|\bar{u}(\boldsymbol{q},\boldsymbol{p})-\bar{u}(\boldsymbol{q}',\boldsymbol{p}')|\leq\epsilon$, note first that

$$|\bar{u}(\boldsymbol{q},\boldsymbol{p})-\bar{u}(\boldsymbol{q}',\boldsymbol{p}')| = \left|\sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p})-1}\pi_{\boldsymbol{p}}(n_{\rm a},\cdot)u_{\boldsymbol{p}}(n_{\rm a})q_{n_{\rm a}} - \sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p}')-1}\pi_{\boldsymbol{p}'}(n_{\rm a},\cdot)u_{\boldsymbol{p}'}(n_{\rm a})q'_{n_{\rm a}}\right|$$

$$\leq \sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p})-1}\left|\pi_{\boldsymbol{p}}(n_{\rm a},\cdot)u_{\boldsymbol{p}}(n_{\rm a})q_{n_{\rm a}}-\pi_{\boldsymbol{p}'}(n_{\rm a},\cdot)u_{\boldsymbol{p}'}(n_{\rm a})q'_{n_{\rm a}}\right| + \sum_{n_{\rm a}=N_{\rm a}(\boldsymbol{p})}^{N_{\rm a}(\boldsymbol{p}')-1}\left|\pi_{\boldsymbol{p}'}(n_{\rm a},\cdot)u_{\boldsymbol{p}'}(n_{\rm a})q'_{n_{\rm a}}\right| \tag{18}$$

We bound each sum in (18) separately. For the first sum, note that

$$\begin{aligned}
&\sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p})-1}\left|\pi_{\boldsymbol{p}}(n_{\rm a},\cdot)u_{\boldsymbol{p}}(n_{\rm a})q_{n_{\rm a}}-\pi_{\boldsymbol{p}'}(n_{\rm a},\cdot)u_{\boldsymbol{p}'}(n_{\rm a})q'_{n_{\rm a}}\right|\\
&\quad=\sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p})-1}\Big|\pi_{\boldsymbol{p}}(n_{\rm a},\cdot)u_{\boldsymbol{p}}(n_{\rm a})(q_{n_{\rm a}}-q'_{n_{\rm a}})+\pi_{\boldsymbol{p}}(n_{\rm a},\cdot)q'_{n_{\rm a}}(u_{\boldsymbol{p}}(n_{\rm a})-u_{\boldsymbol{p}'}(n_{\rm a}))\\
&\qquad\quad+u_{\boldsymbol{p}'}(n_{\rm a})q'_{n_{\rm a}}(\pi_{\boldsymbol{p}}(n_{\rm a},\cdot)-\pi_{\boldsymbol{p}'}(n_{\rm a},\cdot))\Big|\\
&\quad\leq\sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p})-1}U\cdot\left|q_{n_{\rm a}}-q'_{n_{\rm a}}\right|+|u_{\boldsymbol{p}}(n_{\rm a})-u_{\boldsymbol{p}'}(n_{\rm a})|+U\cdot|\pi_{\boldsymbol{p}}(n_{\rm a},\cdot)-\pi_{\boldsymbol{p}'}(n_{\rm a},\cdot)|\\
&\quad\leq\sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p})-1}U\cdot\|\boldsymbol{q}-\boldsymbol{q}'\|+|u_{\boldsymbol{p}}(n_{\rm a})-u_{\boldsymbol{p}'}(n_{\rm a})|+U\cdot(N+1)\|\Pi(\boldsymbol{p})-\Pi(\boldsymbol{p}')\|\\
&\quad\leq\sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p})-1}U\cdot\frac{\epsilon}{4UN}+\frac{\epsilon}{4N}+U\cdot(N+1)\cdot\frac{\epsilon}{4U(N+1)N}=\frac{N_{\rm a}(\boldsymbol{p})}{N}\frac{3\epsilon}{4}\leq\frac{3\epsilon}{4}
\end{aligned}\tag{19}$$

To bound the second sum of (18), simply note that $q_{n_{\rm a}}=0$ for all $n_{\rm a}\geq N_{\rm a}(\boldsymbol{p})$, and so,

$$\begin{aligned}
\sum_{n_{\rm a}=N_{\rm a}(\boldsymbol{p})}^{N_{\rm a}(\boldsymbol{p}')-1}\left|\pi_{\boldsymbol{p}'}(n_{\rm a},\cdot)u_{\boldsymbol{p}'}(n_{\rm a})q'_{n_{\rm a}}\right| &\leq U\sum_{n_{\rm a}=N_{\rm a}(\boldsymbol{p})}^{N_{\rm a}(\boldsymbol{p}')-1}q'_{n_{\rm a}}\\
&=U\sum_{n_{\rm a}=N_{\rm a}(\boldsymbol{p})}^{N_{\rm a}(\boldsymbol{p}')-1}|q'_{n_{\rm a}}-q_{n_{\rm a}}|\\
&=U\sum_{n_{\rm a}=N_{\rm a}(\boldsymbol{p})}^{N_{\rm a}(\boldsymbol{p}')-1}\|\boldsymbol{q}'-\boldsymbol{q}\|\\
&\leq U\cdot(N_{\rm a}(\boldsymbol{p}')-N_{\rm a}(\boldsymbol{p}))\frac{\epsilon}{4U\cdot N}\leq\frac{\epsilon}{4}
\end{aligned}\tag{20}$$

Combining Eq. (18) with (19) and (20), we conclude that $|\bar{u}(\boldsymbol{q},\boldsymbol{p})-\bar{u}(\boldsymbol{q}',\boldsymbol{p}')|\leq\epsilon$. □

Part II - Uniqueness when $N_{\mathrm{a}}(0)=1$

In proving uniqueness for the $N_{\mathrm{a}}(0)=1$ case we make use of the following lemma:

LEMMA 9. *Let $F^{(1)}$ and $F^{(2)}$ be two different distributions such that $F^{(2)} \leq_{\mathrm{st}} F^{(1)}$. Let the random variables $Y^{(j)}$, $j \in \{1,2\}$ represent the stationary number of customers, embedded at arrivals, in a GI/M/1/N queue with service rate μ_{t} and times between potential arrivals distributed according to $F^{(j)}$. Then $Y^{(1)} \leq_{\mathrm{st}} Y^{(2)}$, in the strong sense, i.e., for all $k \in [0, N-1]$, $\mathbf{Pr}\left[Y^{(1)} > k\right] \leq \mathbf{Pr}\left[Y^{(2)} > k\right]$, with strict inequality at some k.*

Proof: For $j \in \{1,2\}$, we refer to the GI/M/1/N system with arrival time distributions $F^{(j)}$ as the j'th system. Let $Y_i^{(j)}$ represent the number of customers in the j'th system, $j \in \{1,2\}$, upon arrival of its i'th customer, $i = 1,2,\ldots$, thus, $Y_i^{(j)}$ converges weakly to $X^{(j)}$ as $i \to \infty$. Assume without loss of generality that $Y_0^{(j)} = 0$. Define the non-negative random variable $D_i^{(j)}$ as the number of potential service completions between the ith and the $i+1$st arrivals, hence it is the number of events in a Poisson process with rate μ_{t} over $F^{(j)}$-distributed random duration. Assuming $F^{(2)} \leq_{\mathrm{st}} F^{(1)}$ implies $D_i^{(2)} \leq_{\mathrm{st}} D_i^{(1)}$ for all i, therefore we can assume without loss of generality that $D_i^{(2)} \leq D_i^{(1)}$ with probability 1 for all i.

A Lindley-like recursion for $Y_i^{(j)}$ is given by

$$Y_{i+1}^{(j)} = [Y_i^{(j)} + \mathbb{1}_{Y_i^{(j)} < N} - D_i^{(j)}]^+,$$

where we note that right-hand-side expression is increasing as a function of $Y_i^{(j)}$ and decreasing as a function of $D_i^{(j)}$. Thus, with $Y_0^{(1)} = Y_0^{(2)} = 0$ and $D_i^{(2)} \leq D_i^{(1)}$, by induction we obtain that $Y_i^{(1)} \leq Y_i^{(2)}$ for all i with probability 1, therefore $Y^{(1)} \leq_{\mathrm{st}} Y^{(2)}$ (in the weak sense). Furthermore, if $F^{(2)}$ and $F^{(1)}$ are not the same distribution then $\mathbf{Pr}\left[D_i^{(2)} = 0\right] > \mathbf{Pr}\left[D_i^{(1)} = 0\right]$. In particular, if $D_i^{(2)} = 0$ and $D_i^{(1)} > 0$ for some i then $Y_{i+1}^{(1)} < Y_{i+1}^{(2)}$, thus,

$$\mathbf{Pr}\left[Y_i^{(1)} < Y_i^{(2)} \quad \text{i.o.}\right] \geq \mathbf{Pr}\left[D_i^{(2)} = 0, D_i^{(1)} > 0\right] > 0,$$

and therefore $\mathbf{Pr}\left[Y^{(1)} < Y^{(2)}\right] > 0$, implying that $Y^{(1)} \leq_{\mathrm{st}} Y^{(2)}$ in the strong sense. □

Suppose now that $N_{\mathrm{a}}(0)=1$, so that customers never join the system when the admission server is busy. Assume customers join an empty admission queue with probability τ. Hence, the admission queue is an M/M/1/1 loss system with potential arrival rate $\tau\lambda$ and service rate μ_{a}. The departure process from the admission queue, which determines the potential arrivals to the treatment queue, is renewal, with inter-event times distributed as the sum of two independent exponential random variables with parameters $\tau\lambda$ and μ_{a}. Thus, the treatment queue is a GI/M/1/N queue with expected time between potential arrivals of $1/(\tau\lambda) + 1/\mu_{\mathrm{a}}$.

Assume the system is stationary and let Y_τ be the (random) number of other customers in the treatment queue at moments of departure from the admission queue. Consider an arbitrary customer arriving at the system. If the admission server is busy, the customer leaves. Otherwise, they join with probability τ, in which case they spend, in expectation, $1/\mu_{\rm a}$ units of time in admission, and then transition to the treatment queue if (and only if) the number of customers in it is less than N, where $N = \lfloor R\mu_{\rm t}/C \rfloor$. The customer is indifferent between joining and balking if and only if their expected utility from joining is zero, namely, if

$$\frac{-C}{\mu_{\rm a}} + \mathbb{E}\left[R - \frac{C(Y_\tau + 1)}{\mu_{\rm t}}\right]^+ = 0. \tag{21}$$

Note that the time between potential arrivals to the treatment queue increases stochastically in τ, therefore, by Lemma 9, Y_τ is stochastically decreasing in τ (in the strong sense). Furthermore, $[R - C(y + 1)/\mu_{\rm t}]^+$ is strictly decreasing in $y \in [0, N-1]$, therefore, the left-hand side of Equation (21) is strictly decreasing in τ. Thus, if there exists some $\tau > 0$ such that equality in (21) holds, then it is unique, otherwise the unique equilibrium is $\tau = 1$.

Proof of Lemma 3

In the proof, we make use of the fact that for any $(n_{\rm a}, n_{\rm t})$, $n_{\rm a} < N_{\rm a}(n_{\rm t})$ if and only if $n_{\rm t} < N_{\rm t}(n_{\rm a})$. This is because $n_{\rm a} < N_{\rm a}(n_{\rm t})$ implies $u(n_{\rm a}, n_{\rm t}) > 0$ which in turn implies $n_{\rm t} < N_{\rm t}(n_{\rm a})$ and vice versa. We also use the fact that for all $n_{\rm t} > 0$, $N_{\rm a}(n_{\rm t}+1) \in \{N_{\rm a}(n_{\rm t}), N_{\rm a}(n_{\rm t}) - 1\}$. This follows from Proposition 1 which implies $u(N_{\rm a}(n_{\rm t}), n_{\rm t}+1) \le u(N_{\rm a}(n_{\rm t}), n_{\rm t}) \le 0$ and $u(N_{\rm a}(n_{\rm t}) - 2, n_{\rm t}+1) \ge u(N_{\rm a}(n_{\rm t}) - 1, n_{\rm t}) > 0$. Following Theorem 4, we can compute the limiting stationary probabilities based on the limits evolving from Equation (3) as $\lambda \to \infty$. To this aim, we first prove two preliminary lemmas:

LEMMA 10. *In the fully observable model, if* $n_{\rm a}, n_{\rm t}$ *are such that* $n_{\rm a} < N_{\rm a}(n_{\rm t})$*, then* $\pi(n_{\rm a}, n_{\rm t}) = O(1/\lambda)$

Proof: Suppose that $n_{\rm a} < N_{\rm a}(n_{\rm t})$, thus $n_{\rm t} < N_{\rm t}(n_{\rm a})$. Consider first the case $n_{\rm a} = 0$. From Equation (3), by substituting $n_{\rm a} = 0$, we obtain, for any $n_{\rm t} < N_{\rm t}(0)$,

$$\pi(0, n_{\rm t}) = \frac{\mu_{\rm a}\pi(1, n_{\rm t}-1)\mathbb{1}_{\{n_{\rm t}>0\}} + \mu_{\rm t}\pi(0, n_{\rm t}+1)}{\mu_{\rm t}\mathbb{1}_{\{n_{\rm t}>0\}} + \lambda} \le \frac{\mu_{\rm a} + \mu_{\rm t}}{\lambda}.$$

For $n_{\rm a} \ge 0$ with $n_{\rm t} < N_{\rm t}(n_{\rm a})$ (equivalently, $n_{\rm a} < N_{\rm a}(n_{\rm t})$) we obtain

$$\begin{aligned}\pi(n_{\rm a}, n_{\rm t}) &= \frac{\lambda\pi(n_{\rm a}-1, n_{\rm t}) + \mu_{\rm a}\pi(n_{\rm a}+1, n_{\rm t}-1)\mathbb{1}_{\{n_{\rm t}>0, n_{\rm a}<N_{\rm a}(n_{\rm t}-1)\}} + \mu_{\rm t}\pi(n_{\rm a}, n_{\rm t}+1)}{\mu_{\rm t}\mathbb{1}_{\{n_{\rm t}>0\}} + \mu_{\rm a}\mathbb{1}_{\{n_{\rm a}>0\}} + \lambda} \\ &\le \frac{\lambda\pi(n_{\rm a}-1, n_{\rm t}) + \mu_{\rm a} + \mu_{\rm t}}{\lambda} = \pi(n_{\rm a}-1, n_{\rm t}) + \frac{\mu_{\rm a} + \mu_{\rm t}}{\lambda} \le (n_{\rm a}+1)\frac{\mu_{\rm a} + \mu_{\rm t}}{\lambda} = O(1/\lambda),\end{aligned} \tag{22}$$

where the second to last transition follows by induction on $n_{\rm a}$. □

LEMMA 11. *In the fully observable model, if* $n_{\rm a}, n_{\rm t}$ *are such that* $n_{\rm a} < N_{\rm a}(n_{\rm t}) - 1$*, then* $\lambda\pi(n_{\rm a}, n_{\rm t}) = O(1/\lambda)$.

Proof: Assume $n_{\mathrm{a}} < N_{\mathrm{a}}(n_{\mathrm{t}}) - 1$. Note that N_{a} is decreasing and satisfies $N_{\mathrm{a}}(n_{\mathrm{t}}) - 1 \le N_{\mathrm{a}}(n_{\mathrm{t}} + 1)$, hence $n_{\mathrm{a}} < N_{\mathrm{a}}(n_{\mathrm{t}} + 1)$, and also $n_{\mathrm{a}} + 1 < N_{\mathrm{a}}(n_{\mathrm{t}}) \le N_{\mathrm{a}}(n_{\mathrm{t}} - 1)$. Based on Lemma 10, this implies that $\pi(n_{\mathrm{a}}, n_{\mathrm{t}}), \pi(n_{\mathrm{a}} + 1, n_{\mathrm{t}} - 1)$ and $\pi(n_{\mathrm{a}}, n_{\mathrm{t}} + 1)$ are all $O(1/\lambda)$. Hence, by Equation (3), we can write

$$\lambda\pi(n_{\mathrm{a}}, n_{\mathrm{t}}) = \lambda\pi(n_{\mathrm{a}} - 1, n_{\mathrm{t}})\mathbb{1}_{\{n_{\mathrm{a}}>0\}} + O(1/\lambda),$$

Replacing n_{a} by $n_{\mathrm{a}} - 1$ and substituting in recursion we get $\lambda\pi(n_{\mathrm{a}}, n_{\mathrm{t}}) = \lambda\pi(0, n_{\mathrm{t}}) + O(1/\lambda)$, and as

$$\lambda\pi(0, n_{\mathrm{t}}) = \mu_{\mathrm{a}}\pi(1, n_{\mathrm{t}} - 1)\mathbb{1}_{\{n_{\mathrm{t}}>0\}} + \mu_{\mathrm{t}}\pi(0, n_{\mathrm{t}} + 1) - \mu_{\mathrm{t}}\pi(0, n_{\mathrm{t}})\mathbb{1}_{\{n_{\mathrm{t}}>0\}} = O(1/\lambda),$$

we conclude that $\lambda\pi(n_{\mathrm{a}}, n_{\mathrm{t}}) = O(1/\lambda)$ as well. □

Assume from now on that $n_{\mathrm{a}} = N_{\mathrm{a}}(n_{\mathrm{t}})$ (otherwise, from Lemma 10, $\lim_{\lambda\to\infty}\pi(n_{\mathrm{a}}, n_{\mathrm{t}}) = 0$), and denote $\lim_{\lambda\to\infty}\pi(N_{\mathrm{a}}(n_{\mathrm{t}}), n_{\mathrm{t}}) = \pi^*_{n_{\mathrm{t}}}$. We next show that the sequence of (limiting) probabilities, $\pi^*_0, \ldots, \pi^*_{N_{\mathrm{t}}(0)}$, which correspond to states $(N_{\mathrm{a}}(0), 0), \ldots, (N_{\mathrm{a}}(N_{\mathrm{t}}(0)), N_{\mathrm{t}}(0))$, satisfy the flow-balance equations of an M/M/1/$N_{\mathrm{t}}(0)$ queue. That is, we show that $\mu_{\mathrm{a}}\pi^*_0 = \mu_{\mathrm{t}}\pi^*_1$, $\mu_{\mathrm{a}}\pi^*_{N_{\mathrm{t}}(0)-1} = \mu_{\mathrm{t}}\pi^*_{N_{\mathrm{t}}(0)}$ and $(\mu_{\mathrm{t}} + \mu_{\mathrm{a}})\pi^*_{n_{\mathrm{t}}} = \mu_{\mathrm{a}}\pi^*_{n_{\mathrm{t}}-1} + \mu_{\mathrm{t}}\pi_{n_{\mathrm{t}}+1}$ for all $n_{\mathrm{t}} \in [1, N_{\mathrm{t}}(0) - 1]$.

Consider first $n_{\mathrm{t}} = 0$ with $n_{\mathrm{a}} = N_{\mathrm{a}}(0)$. If $N_{\mathrm{a}}(1) = N_{\mathrm{a}}(0) - 1$ then Equation (3) gives

$$\mu_{\mathrm{a}}\pi(N_{\mathrm{a}}(0), 0) = \lambda\pi(N_{\mathrm{a}}(0) - 1, 0),$$

and, with Lemmas 10 and 11 also

$$\begin{aligned}\lambda\pi(N_{\mathrm{a}}(0) - 1, 0) &= \lambda\pi(N_{\mathrm{a}}(0) - 2, 0) + \mu_{\mathrm{t}}\pi(N_{\mathrm{a}}(0) - 1, 1) - \mu_{\mathrm{a}}\pi(N_{\mathrm{a}}(0) - 1, 0)\\ &= \mu_{\mathrm{t}}\pi(N_{\mathrm{a}}(1), 1) + O(1/\lambda).\end{aligned}$$

Hence, $\mu_{\mathrm{a}}\pi^*_0 = \lim_{\lambda\to\infty}\mu_{\mathrm{a}}\pi(N_{\mathrm{a}}(0), 0) = \lim_{\lambda\to\infty}\mu_{\mathrm{t}}\pi(N_{\mathrm{a}}(1), 1) = \mu_{\mathrm{t}}\pi^*_1$. If, on the contrary, $N_{\mathrm{a}}(1) = N_{\mathrm{a}}(0)$, then Equation (3) gives

$$\mu_{\mathrm{a}}\pi(N_{\mathrm{a}}(0), 0) = \lambda\pi(N_{\mathrm{a}}(0) - 1, 0) + \mu_{\mathrm{t}}\pi(N_{\mathrm{a}}(0), 1) = \lambda\pi(N_{\mathrm{a}}(0) - 1, 0) + \mu_{\mathrm{t}}\pi(N_{\mathrm{a}}(1), 1),$$

where, by Lemmas 10 and 11,

$$\begin{aligned}\lambda\pi(N_{\mathrm{a}}(0) - 1, 0) &= \lambda\pi(N_{\mathrm{a}}(0) - 2, 0)\mathbb{1}_{\{N_{\mathrm{a}}(0)>2\}} + \mu_{\mathrm{t}}\pi(N_{\mathrm{a}}(0) - 1, 1) - \mu_{\mathrm{a}}\pi(N_{\mathrm{a}}(0) - 1, 0)\\ &= O(1/\lambda).\end{aligned}$$

therefore $\mu_{\mathrm{a}}\pi^*_0 = \mu_{\mathrm{t}}\pi^*_1$.

Consider now the case $n_{\mathrm{t}} = N_{\mathrm{t}}(0)$. Thus $N_{\mathrm{a}}(n_{\mathrm{t}}) = 0$, for which it must hold that $N_{\mathrm{a}}(n_{\mathrm{t}} - 1) = 1$ and Equation (3) together with Lemma 11 give

$$\mu_{\mathrm{t}}\pi(N_{\mathrm{a}}(n_{\mathrm{t}}), n_{\mathrm{t}}) = \mu_{\mathrm{a}}\pi(N_{\mathrm{a}}(n_{\mathrm{t}}) + 1, n_{\mathrm{t}} - 1) + O(1/\lambda) = \mu_{\mathrm{a}}\pi(N_{\mathrm{a}}(n_{\mathrm{t}} - 1), n_{\mathrm{t}} - 1) + O(1/\lambda),$$

therefore $\mu_{\mathrm{a}}\pi^*_{N_{\mathrm{t}}(0)-1} = \mu_{\mathrm{t}}\pi^*_{N_{\mathrm{t}}(0)}$

Lastly, we prove the claim assuming $n_{\rm t} > 0$ and $n_{\rm a} = N_{\rm a}(n_{\rm t}) > 0$. We show that $(\mu_{\rm t} + \mu_{\rm a})\pi^*_{n_{\rm t}} = \mu_{\rm a}\pi^*_{n_{\rm t}-1} + \mu_{\rm t}\pi_{n_{\rm t}+1}$. Recall that $N_{\rm a}(n_{\rm t}+1) \in \{N_{\rm a}(n_{\rm t}), N_{\rm a}(n_{\rm t})-1\}$ for any $n_{\rm t}$, thus, we distinguish between 4 cases:

Case I: Suppose $N_{\rm a}(n_{\rm t}+1) = N_{\rm a}(n_{\rm t})$ and $N_{\rm a}(n_{\rm t}-1) = N_{\rm a}(n_{\rm t})+1$, then, by Lemma 10 ,

$$\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}+1) = \pi(N_{\rm a}(n_{\rm t}+1)-1, n_{\rm t}+1) = O(1/\lambda),$$

as well as

$$\pi(N_{\rm a}(n_{\rm t}), n_{\rm t}-1) = \pi(N_{\rm a}(n_{\rm t}-1)-1, n_{\rm t}-1) = O(1/\lambda),$$

and from Equation (3) for state $(N_{\rm a}(n_{\rm t})-1, n_{\rm t})$ we have $\lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}) = O(1/\lambda)$. Hence, using Equation (3) for state $(N_{\rm a}(n_{\rm t}), n_{\rm t})$ we get

$$\begin{aligned}(\mu_{\rm t}+\mu_{\rm a})\,\pi(N_{\rm a}(n_{\rm t}), n_{\rm t}) &= \lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}) + \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t})+1, n_{\rm t}-1) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}), n_{\rm t}+1)\\ &= \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t}-1), n_{\rm t}-1) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}+1), n_{\rm t}+1) + O(1/\lambda).\end{aligned}$$

Case II: Suppose $N_{\rm a}(n_{\rm t}+1) = N_{\rm a}(n_{\rm t})$ and $N_{\rm a}(n_{\rm t}-1) = N_{\rm a}(n_{\rm t})$, thus, Equation (3) for state $(N_{\rm a}(n_{\rm t})-1, n_{\rm t})$ shows that

$$\lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}) = \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t})-1+1, n_{\rm t}-1) + O(1/\lambda) = \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t}-1), n_{\rm t}-1) + O(1/\lambda).$$

Hence, using Equation (3) for $(N_{\rm a}(n_{\rm t}), n_{\rm t})$ with $N_{\rm a}(n_{\rm t}) = N_{\rm a}(n_{\rm t}-1)$ we get

$$\begin{aligned}(\mu_{\rm t}+\mu_{\rm a})\,\pi(N_{\rm a}(n_{\rm t}), n_{\rm t}) &= \lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}), n_{\rm t}+1)\\ &= \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t}-1), n_{\rm t}-1) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}+1), n_{\rm t}+1) + O(1/\lambda).\end{aligned}$$

Case III: Suppose $N_{\rm a}(n_{\rm t}+1) = N_{\rm a}(n_{\rm t})-1$ and $N_{\rm a}(n_{\rm t}-1) = N_{\rm a}(n_{\rm t})+1$, thus, Equation (3) for state $(N_{\rm a}(n_{\rm t})-1, n_{\rm t})$ shows that

$$\lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}) = \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}+1) + O(1/\lambda) = \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}+1), n_{\rm t}+1) + O(1/\lambda).$$

Hence, using Equation (3) for $(N_{\rm a}(n_{\rm t}), n_{\rm t})$ with $N_{\rm a}(n_{\rm t}) > N_{\rm a}(n_{\rm t}+1)$ we get

$$\begin{aligned}(\mu_{\rm t}+\mu_{\rm a})\,\pi(N_{\rm a}(n_{\rm t}), n_{\rm t}) &= \lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}) + \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t})+1, n_{\rm t}-1)\\ &= \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t}-1), n_{\rm t}-1) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}+1), n_{\rm t}+1) + O(1/\lambda).\end{aligned}$$

Case IV: Suppose $N_{\rm a}(n_{\rm t}+1) = N_{\rm a}(n_{\rm t})-1$ and $N_{\rm a}(n_{\rm t}-1) = N_{\rm a}(n_{\rm t})$, thus, Equation (3) (for $(N_{\rm a}(n_{\rm t})-1, n_{\rm t})$) shows that

$$\begin{aligned}\lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}) &= \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t})-1+1, n_{\rm t}-1) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t}+1) + O(1/\lambda)\\ &= \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t}-1), n_{\rm t}-1) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}+1), n_{\rm t}+1) + O(1/\lambda).\end{aligned}$$

Hence, using Equation (3) for $(N_{\rm a}(n_{\rm t}), n_{\rm t})$ with $N_{\rm a}(n_{\rm t}) = N_{\rm a}(n_{\rm t}-1)$ and $N_{\rm a}(n_{\rm t}) > N_{\rm a}(n_{\rm t}+1)$ we get

$$\begin{aligned}(\mu_{\rm t}+\mu_{\rm a})\,\pi(N_{\rm a}(n_{\rm t}), n_{\rm t}) &= \lambda\pi(N_{\rm a}(n_{\rm t})-1, n_{\rm t})\\ &= \mu_{\rm a}\pi(N_{\rm a}(n_{\rm t}-1), n_{\rm t}-1) + \mu_{\rm t}\pi(N_{\rm a}(n_{\rm t}+1), n_{\rm t}+1) + O(1/\lambda).\end{aligned}$$

Overall, we conclude that the probabilities $\pi^*_0, \dots \pi^*_{N_{\rm t}(0)}$ satisfy the flow-balance equations of the well-known M/M/1/$N_{\rm t}(0)$ queueing model with arrival and service rates $\mu_{\rm a}$ and $\mu_{\rm t}$ respectively. Based on Kleinrock (1975) Chapter 3.6, we thus have

$$\pi^*_{n_{\rm t}} = \frac{\eta^{n_{\rm t}}(1-\eta)}{1-\eta^{N_{\rm t}(0)+1}},$$

for all $n_{\rm t} \in [0, N_{\rm t}(0)]$, from which the throughput and social-welfare expressions readily follow.

Proof of Lemma 4

We first consider a fixed τ-threshold strategy $\boldsymbol{p}^\tau$ with $p^\tau_0 = 1$, i.e., $\tau \geq 1$, and show in part I that for such a strategy, the admission queue length (weakly) converges to a constant $N_{\rm a}(\boldsymbol{p}^\tau) \geq 1$ and that the stationary treatment-queue length converges to the stationary queue length of an M/M/1/N, from which the limiting throughput and social welfare can be calculated. Afterwards, we show in part II that under the condition $\nu_{\rm t} > C$, the equilibrium strategy in the limit must be a τ-threshold strategy for some $\tau \geq 1$.

Part I

Fix $\tau \geq 1$ with the corresponding τ-threshold strategy $\boldsymbol{p}^\tau$. Consider $n_{\rm a} < N_{\rm a}(\boldsymbol{p}^\tau)$. Similarly to the proof of Lemma 3, from Equation (17), it can be seen that

$$\begin{aligned}
\pi_{\boldsymbol{p}^\tau}(n_{\rm a}, n_{\rm t}) &\leq \frac{\lambda \pi_{\boldsymbol{p}^\tau}(n_{\rm a}-1, n_{\rm t}) \mathbb{1}_{\{n_{\rm a}>0\}} + \mu + 2\mu_{\rm t}}{p^\tau_{n_{\rm a}} \lambda} \\
&\leq \frac{1}{p^\tau_{n_{\rm a}}} \left(\pi_{\boldsymbol{p}^\tau}(n_{\rm a}-1, n_{\rm t}) \mathbb{1}_{\{n_{\rm a}>0\}} + \frac{\mu + 2\mu_{\rm t}}{\lambda} \right) \\
&= \frac{1}{p^\tau_{n_{\rm a}}} \left(\pi_{\boldsymbol{p}^\tau}(n_{\rm a}-2, n_{\rm t}) \mathbb{1}_{\{n_{\rm a}-1>0\}} + \frac{\mu + 2\mu_{\rm t}}{\lambda} + \frac{\mu + 2\mu_{\rm t}}{\lambda} \right) \\
&\leq \frac{1}{p^\tau_{n_{\rm a}}} \cdot n_{\rm a} \frac{\mu + 2\mu_{\rm t}}{\lambda} = O(1/\lambda),
\end{aligned}$$

where the second to last inequality follows recursive substitutions and the fact that $p^\tau_{n_{\rm a}} = 1$ for all $n_{\rm a} < N_{\rm a}(\boldsymbol{p}^\tau) - 1$. It therefore must follow that for all $n_{\rm a} < N_{\rm a}(\boldsymbol{p}^\tau)$, $\pi_{\boldsymbol{p}^\tau}(n_{\rm a}, \cdot) = O(1/\lambda)$, and from $\pi_{\boldsymbol{p}^\tau}(N_{\rm a}(\boldsymbol{p}^\tau), \cdot) = 1 - \sum_{n_{\rm a}=0}^{N_{\rm a}(\boldsymbol{p}^\tau)-1} \pi_{\boldsymbol{p}^\tau}(n_{\rm a}, \cdot)$, it further follows that $\lim_{\lambda\to\infty} \pi_{\boldsymbol{p}^\tau}(N_{\rm a}(\boldsymbol{p}^\tau), \cdot) = 1$. Based on this, it can be seen from Equation (17) that if $n_{\rm a} < N_{\rm a}(\boldsymbol{p}^\tau) - 1$, as we have $\pi_{\boldsymbol{p}^\tau}(n_{\rm a}+1, n_{\rm t}-1) = O(1/\lambda)$, then

$$\lambda p_{n_{\rm a}} \pi_{\boldsymbol{p}}(n_{\rm a}, n_{\rm t}) = \lambda p_{n_{\rm a}-1} \pi_{\boldsymbol{p}}(n_{\rm a}-1, n_{\rm t}) \mathbb{1}_{\{n_{\rm a}>0\}} + O(1/\lambda).$$

For $n_{\rm a} = 0$, as we have $p^\tau_0 = 1$, we get $\lambda \pi_{\boldsymbol{p}^\tau}(0, n_{\rm t}) = O(1/\lambda)$, and therefore $\lambda \pi_{\boldsymbol{p}^\tau}(n_{\rm a}, n_{\rm t}) = O(1/\lambda)$ for all $n_{\rm a} < N_{\rm a}(\boldsymbol{p}^\tau) - 1$ and $n_{\rm t} \leq N$.

Assume from now on that $n_{\rm a} = N_{\rm a}(\boldsymbol{p}^\tau)$ (otherwise $\lim_{\lambda\to\infty} \pi_{\boldsymbol{p}^\tau}(n_{\rm a}, n_{\rm t}) = 0$), and denote $\lim_{\lambda\to\infty} \pi_{\boldsymbol{p}^\tau}(N_{\rm a}(\boldsymbol{p}^\tau), n_{\rm t}) = \pi^*_{n_{\rm t}}$. Similarly to the proof of Lemma 3, we next show that the sequence of (limiting) probabilities, $\pi^*_0, \dots, \pi^*_N$, which correspond to states $(N_{\rm a}(\boldsymbol{p}^\tau), 0), \dots, (N_{\rm a}(\boldsymbol{p}^\tau), N)$, satisfy the flow-balance equations of an M/M/1/N queue.

Consider $n_{\mathrm{t}}=0$, then Equation (17) for state $(N_{\mathrm{a}}(\boldsymbol{p}^\tau),0)$ gives

$$\mu_{\mathrm{a}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),0)=\lambda p^\tau_{N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,0)+\mu_{\mathrm{t}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),1),$$

and for state $(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,0)$ it gives

$$\lambda p^\tau_{N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,0)=\lambda p^\tau_{N_{\mathrm{a}}(\boldsymbol{p}^\tau)-2}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-2,0)+\mu_{\mathrm{t}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,1)=O(1/\lambda),$$

hence $\mu_{\mathrm{a}}\pi^*_0=\mu_{\mathrm{t}}\pi^*_1$.

Consider now $n_{\mathrm{t}}=N$, then Equation (17) for state $(N_{\mathrm{a}}(\boldsymbol{p}^\tau),N)$ gives

$$(\mu_{\mathrm{a}}+\mu_{\mathrm{t}})\pi_{\boldsymbol{p}}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),N)=\lambda p_{N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,N),$$

and for state $(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,N)$ it gives

$$\lambda p^\tau_{N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,N)=\mu_{\mathrm{a}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),N)+\mu_{\mathrm{a}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),N-1)+O(1/\lambda),$$

which implies

$$\mu_{\mathrm{t}}\pi_{\boldsymbol{p}}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),N)=\mu_{\mathrm{a}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),N-1)+O(1/\lambda),$$

hence $\mu_{\mathrm{t}}\pi^*_N=\mu_{\mathrm{a}}\pi^*_{N_{\mathrm{a}}(0)-1}$.

Lastly, consider $n_{\mathrm{t}}\in[1,N-1]$, then Equation (17) for state $(N_{\mathrm{a}}(\boldsymbol{p}^\tau),n_{\mathrm{t}})$ gives

$$(\mu_{\mathrm{a}}+\mu_{\mathrm{t}})\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}),n_{\mathrm{t}})=\lambda p_{N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p})-1,n_{\mathrm{t}})+\mu_{\mathrm{t}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),n_{\mathrm{t}}+1),$$

and for state $(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,n_{\mathrm{t}})$ it gives

$$\lambda p_{N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau)-1,n_{\mathrm{t}})=\mu_{\mathrm{a}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),n_{\mathrm{t}}-1)+O(1/\lambda),$$

which implies

$$(\mu_{\mathrm{a}}+\mu_{\mathrm{t}})\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),n_{\mathrm{t}})=\mu_{\mathrm{a}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),n_{\mathrm{t}}-1)+\mu_{\mathrm{t}}\pi_{\boldsymbol{p}^\tau}(N_{\mathrm{a}}(\boldsymbol{p}^\tau),n_{\mathrm{t}}+1)+O(1/\lambda),$$

hence $(\mu_{\mathrm{a}}+\mu_{\mathrm{t}})\pi^*_{n_{\mathrm{t}}}=\mu_{\mathrm{a}}\pi^*_{n_{\mathrm{t}}-1}+\mu_{\mathrm{t}}\pi^*_{n_{\mathrm{t}}+1}$. Similarly to the proof of Lemma 3, we conclude that

$$\pi^*_{n_{\mathrm{t}}}=\frac{\eta^{n_{\mathrm{t}}}(1-\eta)}{1-\eta^{N+1}},$$

for all $n_{\mathrm{t}}\in[0,N]$, from which the throughput expression readily follows, and the expected treatment-queue length is

$$\sum_{n_{\mathrm{t}}=0}^{N}n_{\mathrm{t}}\pi^*_{n_{\mathrm{t}}}=\frac{\eta}{1-\eta}-\frac{(N+1)\eta^{N+1}}{1-\eta^{N+1}}.$$

Because the admission-queue process is reversible, its effective arrival process is statistically identical to its departure process. Because $\pi_{\boldsymbol{p}^\tau}(0,\cdot)\to 0$ as $\lambda\to\infty$, we deduce that in the limit, the departure process from the admission queue, and hence also the effective arrival process to the system, are both Poisson with rate $\mu_{\rm a}$.

Any customer, upon transitioning to the treatment queue, observes the time-averaged (stationary) queue length of an M/M/1/N. Hence, their total value from joining the system with admission queue of length $n_{\rm a} < N_{\rm a}(\boldsymbol{p}^\tau)$ can be separated into the cost spent waiting in the admission queue, $C(n_{\rm a}+1)/\mu_{\rm a}$, and the expected utility from transitioning to the treatment queue, namely,

$$\lim_{\lambda\to\infty} u_{\boldsymbol{p}^\tau}(n_{\rm a}) = -\frac{C(n_{\rm a}+1)}{\mu_{\rm a}} + (1-\pi^*_N)\left(R - C\frac{\sum_{n_{\rm t}=0}^{N-1} n_{\rm t}\pi^*_{n_{\rm t}}}{1-\pi^*_N} - \frac{C}{\mu_{\rm t}}\right) = -\frac{C(n_{\rm a}+1)}{\mu_{\rm a}} + \frac{\nu_{\rm t}}{\mu_{\rm a}}, \tag{23}$$

where the equality follows substitution and basic algebra. In addition, $\pi_{\boldsymbol{p}^\tau}(N_{\rm a}(\boldsymbol{p}^\tau),\cdot)\to 1$, which means that the aggregate social cost inflicted per unit of time by customers waiting in admission is $CN_{\rm a}(\boldsymbol{p}^\tau)$. Consequently, the corresponding limiting social welfare can be computed as

$$\mu_{\rm a}\frac{\nu_{\rm t}}{\mu_{\rm a}} - CN_{\rm a}(\boldsymbol{p}) = \nu_{\rm t} - CN_{\rm a}(\boldsymbol{p}) = \mu_{\rm a}\lim_{\lambda\to\infty} u_{\boldsymbol{p}^\tau}(N_{\rm a}(\boldsymbol{p}^\tau)-1).$$

Part II

We first prove the following auxiliary lemma:

LEMMA 12. *Suppose $\nu_t > C$, then in equilibrium $\tau \geq 1$.*

Proof: Consider a fixed equilibrium threshold τ, and suppose for the sake of contradiction that $\tau < 1$. Then, as explained in part II of the proof of Theorem 2, the treatment queue is a GI/M/1/N system, where a typical interarrival time duration is distributed as the sum of two exponential random variables with parameters $\mu_{\rm a}$ and $\tau\lambda$. Note that this interarrival-time distribution is stochastically decreasing in λ, hence by Lemma 9 the stationary observed length at the treatment queue increases in λ. Also note that as $\lambda\to\infty$, the arrival process to the treatment queue approaches Poisson with rate $\mu_{\rm a}$. Hence, the utility from joining the system when the admission queue is empty is decreasing in λ, and similarly to part I, approaches

$$-\frac{C}{\mu_{\rm a}} + \frac{\nu_{\rm t}}{\mu_{\rm a}} > 0$$

as $\lambda\to\infty$. This implies that the utility from joining an empty admission queue is strictly positive for any λ, therefore in equilibrium it must follow that $p_0^\tau = 1$, in contradiction to $\tau < 1$. □

Suppose $\nu_t > C$. Thus, by Lemma 12 it must follow that in equilibrium, $\tau \geq 1$ for all $\lambda > 0$. By part I, the limiting throughput in this case is given by that of an M/M/1/N queue. We therefore focus in this part on Equation (9) for the limiting social welfare.

Let $\lambda_1,\lambda_2,\ldots$ be an increasing and unbounded sequence, and let $\tau_1,\tau_2,\ldots$ be a sequence of corresponding equilibrium thresholds, such that the τ_n-threshold strategy $\boldsymbol{p}^{\tau_n}$ is an equilibrium corresponding with arrival rate λ_n. Note that from Lemma 12 and the definition of $N_{\rm a}(0)$, the sequence $\tau_1,\tau_2,\ldots$ is bounded in $[1, N_{\rm a}(0))$, thus we can assume without loss of generality that it is convergent (otherwise we choose a convergent subsequence of it), and we denote its limit by τ'. Our goal is to show that an accumulation point τ' of $\{\tau_n\}$ must satisfy $\lfloor \nu_{\rm t}/C \rfloor - 1 < \tau' \leq \lfloor \nu_{\rm t}/C \rfloor$. If $\boldsymbol{p}^{\tau'}$ is the corresponding τ'-threshold strategy, then the latter inequality implies $N_{\rm a}(\boldsymbol{p}^{\tau'}) = \lfloor \nu_{\rm t}/C \rfloor$, and the social welfare expression in Equation (9) follows immediately from part I. We begin by showing that $\tau' \leq \lfloor \nu_{\rm t}/C \rfloor$.

Assume, for the sake of contradiction that $\tau' > \lfloor \nu_{\rm t}/C \rfloor$, thus, $N_{\rm a}(\boldsymbol{p}^{\tau'}) \geq \lfloor \nu_{\rm t}/C \rfloor + 1$. Let $\epsilon = \tau' - \lfloor \tau' \rfloor$ if $\tau' > \lfloor \tau' \rfloor$ and 1 otherwise, thus $\epsilon > 0$. Hence, for all $\tau > \tau' - \epsilon$, $N_{\rm a}(\boldsymbol{p}^{\tau}) \geq N_{\rm a}(\boldsymbol{p}^{\tau'}) \geq \lfloor \nu_{\rm t}/C \rfloor + 1$. Define

$$\epsilon_u = C\left(\left\lfloor \frac{\nu_{\rm t}}{C} \right\rfloor + 1\right) - \nu_{\rm t} > 0.$$

The assumption that $\{\tau_n\}$ converges to τ' together with part I of the proof (Equation (23)) imply that there exists some sufficiently large n for which $\tau_n > \tau' - \epsilon$, and in addition, λ_n is sufficiently large such that

$$u_{\boldsymbol{p}^{\tau_n}}(N_{\rm a}(\boldsymbol{p}^{\tau_n}) - 1)\,|_{\lambda=\lambda_n} < \lim_{\lambda\to\infty} u_{\boldsymbol{p}^{\tau_n}}(N_{\rm a}(\boldsymbol{p}^{\tau_n}) - 1) + \frac{\epsilon_u}{\mu_{\rm a}}.$$

Thus,

$$u_{\boldsymbol{p}^{\tau_n}}(N_{\rm a}(\boldsymbol{p}^{\tau_n}) - 1)\,|_{\lambda=\lambda_n} < \frac{\nu_{\rm t} - CN_{\rm a}(\boldsymbol{p}^{\tau_n})}{\mu_{\rm a}} + \frac{\epsilon_u}{\mu_{\rm a}} \leq \frac{\nu_{\rm t} - C(\lfloor \nu_{\rm t}/C \rfloor + 1) + \epsilon_u}{\mu_{\rm a}} = 0.$$

Hence, with arrival rate λ_n, the best response against the τ_n-threshold strategy is to balk when observing $N_{\rm a}(\boldsymbol{p}^{\tau_n}) - 1 < \tau_n$ in admission, which is in contradiction to τ_n being the equilibrium threshold. It therefore follows that $\tau' \leq \lfloor \nu_{\rm t}/C \rfloor$.

We next show that $\tau' > \lfloor \nu_{\rm t}/C \rfloor - 1$. Assume, again for the sake of contradiction, that $\tau' \leq \lfloor \nu_{\rm t}/C \rfloor - 1$, thus, $N_{\rm a}(\boldsymbol{p}^{\tau'}) \leq \lfloor \nu_{\rm t}/C \rfloor - 1$. Let $\epsilon = 1 - (\tau' - \lfloor \tau' \rfloor) > 0$, so that for all $\tau < \tau' + \epsilon$, $N_{\rm a}(\boldsymbol{p}^{\tau}) \leq N_{\rm a}(\boldsymbol{p}^{\tau'})$. Assuming $\nu_{\rm t}/C$ is non-integer, we redefine ϵ_u as

$$\epsilon_u = \nu_{\rm t} - C\lfloor \nu_{\rm t}/C \rfloor > 0.$$

Thus, following part I (Equation (23)), there exists some sufficiently large n such that $\tau_n < \tau' + \epsilon$, and in addition, λ_n is large enough so that

$$u_{\boldsymbol{p}^{\tau_n}}(N_{\rm a}(\boldsymbol{p}^{\tau_n}) - 1) - u_{\boldsymbol{p}^{\tau_n}}(N_{\rm a}(\boldsymbol{p}^{\tau_n})) < \frac{C + \epsilon_u}{\mu_{\rm a}}.$$

This implies

$$u_{\boldsymbol{p}^{\tau_n}}(N_{\rm a}(\boldsymbol{p}^{\tau_n})) > u_{\boldsymbol{p}^{\tau_n}}(N_{\rm a}(\boldsymbol{p}^{\tau_n}) - 1) - \frac{C + \epsilon_u}{\mu_{\rm a}} \geq \frac{\nu_{\rm t} - C(\lfloor \nu_{\rm t}/C \rfloor - 1) - C - \epsilon_u}{\mu_{\rm a}} \geq 0,$$

while we have, from the definition of $N_{\rm a}(\boldsymbol{p}^{\tau_n})$, that $p^{\tau_n}_{N_{\rm a}(\boldsymbol{p}^{\tau_n})} = 0$, in contradiction to the equilibrium criterion. We therefore conclude that $\lfloor \nu_{\rm t}/C \rfloor - 1 < \tau' \leq \lfloor \nu_{\rm t}/C \rfloor$.

Equilibrium approximation scheme in the partially observable model

We calculate the equilibrium numerically using an adaptation of the (deterministic) fixed point iteration introduced in Ravner and Snitkovsky (2023). Recall from the proof of Theorem 2 the definition of $\mathcal{S} = [0,1]^{N_\mathrm{a}(0)}$, the strategy space, which is compact and convex, and let $\mathbf{\Pi}_{\mathcal{S}} : \mathbb{R}^{N_\mathrm{a}(0)} \to \mathcal{S}$ denote the projection operator onto $\mathcal{S}$. Note that because $\mathcal{S}$ is the Cartesian product of unit intervals, thus, for any $\boldsymbol{v} \in \mathbb{R}^{N_\mathrm{a}(0)}$, $\mathbf{\Pi}_{\mathcal{S}}(\boldsymbol{v})$ simply confines each coordinate of $\boldsymbol{v}$ to $[0,1]$. Next we introduce the utility vector function $\boldsymbol{u} : \mathcal{S} \to \mathbb{R}^{N_\mathrm{a}(0)}$. Recall from Equation (4) the definition of $u_{\boldsymbol{p}}(n_\mathrm{a})$. Then, for any $\boldsymbol{p} \in \mathcal{S}$, define $\boldsymbol{u}(\boldsymbol{p}) = (u_0(\boldsymbol{p}), \ldots, u_{N_\mathrm{a}(0)}(\boldsymbol{p}))$ such that for any $n_\mathrm{a} = 0, 1, \ldots N_\mathrm{a}(0)$,

$$u_{n_\mathrm{a}}(\boldsymbol{p}) = \begin{cases} u_{\boldsymbol{p}}(n_\mathrm{a}) & \text{if } n_\mathrm{a} \le N_\mathrm{a}(\boldsymbol{p}), \\ 0 & \text{otherwise.} \end{cases}$$

Let $\gamma_1, \gamma_2, \ldots$ be a sequence of positive real values, referred to as the step-size sequence, and assume it is non-summable and diminishing to 0, namely, that $\sum_{n=1}^{\infty} \gamma_n = \infty$ and $\gamma_n \to 0$. Starting with an initial guess $\boldsymbol{p}^{(0)}$, at any iteration $n \ge 1$, we calculate the n'th iterate as

$$\boldsymbol{p}^{(n)} = \mathbf{\Pi}_{\mathcal{S}} \left(\boldsymbol{p}^{(n-1)} + \gamma_n \boldsymbol{u}(\boldsymbol{p}^{(n-1)}) \right).$$

From Ravner and Snitkovsky (2023), it is known that if this iterative scheme converges, then its limit $\boldsymbol{p}^*$ must be an equilibrium strategy, that is, $\boldsymbol{p}^*$ satisfies $\boldsymbol{p}^* \in \mathcal{BR}(\boldsymbol{p}^*)$ (see Equation (16)).

In the numeric calculations we set $\gamma_n = 1/(R\sqrt{n})$. For the initial strategy, we typically set $\boldsymbol{p}^{(0)}$ as the τ-threshold strategy with $\tau = 1$, though by experimenting with different initial conditions we observe that the iterative scheme seems to converge to a limit independent of the initial $\boldsymbol{p}^{(0)}$. This is a strong evidence that the conditions for convergence of the iterative scheme are satisfied in our framework. We terminate the algorithm at iteration n either if $\boldsymbol{p}^{(n)}$ satisfies an ϵ-approximate equilibrium condition, i.e., if, for some arbitrary $\boldsymbol{q} \in \mathcal{BR}(\boldsymbol{p}^{(n)})$

$$\left| \sum_{n_\mathrm{a}=0}^{N_\mathrm{a}(\boldsymbol{p}^{(n)})-1} \pi_{\boldsymbol{p}^{(n)}}(n_\mathrm{a}, \cdot) u_{\boldsymbol{p}^{(n)}}(n_\mathrm{a}) \cdot (p^{(n)}_{n_\mathrm{a}} - q_{n_\mathrm{a}}) \right| < \epsilon,$$

where we typically set $\epsilon = R \cdot 10^{-4}$. This inequality implies that if all customers join according to $\boldsymbol{p}^{(n)}$, a customer cannot improve their ex-ante utility by more than ϵ if they deviate from the strategy $\boldsymbol{p}^{(n)}$ to some other strategy, in particular, to a strategy that is a best response to $\boldsymbol{p}^{(n)}$. Hence, the left-hand side of the inequality yields the same value for any $\boldsymbol{q} \in \mathcal{BR}(\boldsymbol{p}^{(n)})$. It is worth mentioning that if $\boldsymbol{p}^{(n)} = \boldsymbol{p}^{(n-1)}$, then $\boldsymbol{p}^{(n)}$ must satisfy the equilibrium condition $\boldsymbol{p}^{(n)} \in \mathcal{BR}(\boldsymbol{p}^{(n)})$, and therefore also the ϵ-approximate equilibrium condition for any $\epsilon > 0$.