RESEARCH ARTICLE

Flattening the curve: Insights from queueing theory

Sergio Palomo\textsuperscript{1}, Jamol J. Pender\textsuperscript{1,2,*}, William A. Massey\textsuperscript{3}, Robert C. Hampshire\textsuperscript{4}

\textsuperscript{1}Cornell University Systems Engineering, Ithaca, NY, United States of America, \textsuperscript{2}Cornell University Operations Research and Information Engineering, Ithaca, NY, United States of America, \textsuperscript{3}Princeton University, Operations Research and Financial Engineering, Princeton, NJ, United States of America, \textsuperscript{4}University of Michigan, Gerald Ford School of Public Policy, Ann Arbor, MI, United States of America

* jjp274@cornell.edu

Abstract

The worldwide outbreak of the coronavirus was first identified in 2019 in Wuhan, China. Since then, the disease has spread worldwide. As it is currently spreading in the United States, policy makers, public health officials and citizens are racing to understand the impact of this virus on the United States healthcare system. They fear a rapid influx of patients overwhelming the healthcare system and leading to unnecessary fatalities. Most countries and states in America have introduced mitigation strategies, such as using social distancing to decrease the rate of newly infected people. This is what is usually meant by flattening the curve. In this paper, we use queueing theoretic methods to analyze the time evolution of the number of people hospitalized due to the coronavirus. Given that the rate of new infections varies over time as the pandemic evolves, we model the number of coronavirus patients as a dynamical system based on the theory of infinite server queues with time inhomogeneous Poisson arrival rates. With this model we are able to quantify how flattening the curve affects the peak demand for hospital resources. This allows us to characterize how aggressive societal policy needs to be to avoid overwhelming the capacity of healthcare system. We also demonstrate how curve flattening impacts the elapsed lag between the times of the peak rate of hospitalizations and the peak demand for the hospital resources. Finally, we present empirical evidence from Italy and the United States that supports the insights from our model analysis.

1 Introduction

In December of 2019, the coronavirus was first identified in Wuhan, China. This is a city of over 11 million people in the Hubei province of China. COVID-19, the disease caused by the coronavirus, grew by several thousand per day in China between late January 2020 and early February 2020, the first local peak of their epidemic. The number of daily infections decreased significantly due to stringent containment efforts.

As the viral spread became a global pandemic, large outbreaks in South Korea, Iran, Italy, Spain, and the United States all sparked interest in coronavirus research. There were 219 million cases and 4,510,124 deaths confirmed worldwide, as of September 1\textsuperscript{st} 2021. The World
Health Organization then officially named COVID-19 a global pandemic. The rapid spread of the virus during that time then led to over 200 countries reporting cases.

Lawmakers and public health experts such as Dr. Anthony Fauci, National Institute of Allergy and Infectious Diseases director, encouraged “social distancing” as a measure to flatten the curve. This implies that we want to reduce the daily peak number of infected people and spread it out over time. Many people understand how flattening the curve impacts the healthcare system from an empirical or qualitative perspective, see [1]. However, there needs to be more quantitative research in this area and relatively more research on how much reduction is needed to flatten the curve.

Everyday the hospital has a demand for a finite but reusable set of shared resources that include beds, ventilators, and medical staff. Our paper models this resource demand as an infinite server queue with a non-homogeneous Poisson arrival process and random, mutually independent service times having some common service time distribution. Such a non-stationary infinite server model affords us a great deal of modeling flexibility. Examples of what the arrival process can model include

- The daily number of infected people.
- The daily number of hospitalizations.

Moreover, the service times can model, respectively

- The time of being infected.
- The time the patient spends in the hospital.

These infinite server queueing models also describe more realistic complex behaviors such as multi-modal patient arrival surges. We can reduce them to the superposition of queues with simple uni-modal mean arrival rate functions. We illustrate this type of peak behaviour by using mean arrival rate examples involving scaled Gaussian and the scaled gamma density functions. Here, we obtain explicit expressions for the peak mean demand rate of the queueing system as well as the time of this peak rate.

Our first goal is to show the global community how the operations research field of non-stationary queueing theory helps us understand the impact of COVID-19 on healthcare systems. Our second goal is to show how leveraging these insights from queues with time-varying rates leads to simple descriptions of COVID-19 pandemic-like hospital patient dynamics. More specifically, we show how to calculate the following metrics:

- The time of the peak infection rate.
- The time of the peak mean number of infected people.
- The time lag between them.

We do a similar analysis for the times of locally peak hospitalization rates and the corresponding times with their peak numbers of patients. Both of these results come from studying the same general infinite server queueing model. We also do this analysis explicitly for the cases of scaled Gaussian and gamma density arrival rates. Thus, by using these non-stationary infinite server models, we can demonstrate the impact of public health policies such as social distancing, mask wearing, and curve flattening.

Operationally, what can we learn from situations already observed in China, Europe, and New York City to avoid similar outcomes in other places? From the data, we see that “every hotspot has its own curve”, as shown in Hilk et al [2]. One important advantage of our work is that it can be used to study each individual curve for individual regions around the world. The
geographic and temporal clustering of outbreaks have the potential to create a health care system collapse. Many states in the United States have imposed “stay at home” orders to contain the spread of the disease. However, since many people are asymptomatic, we can show through the analysis of our queueing examples that containment measures may take weeks to have an observable effect in the infection data.

Although hospital bed capacity is an important concern for health care officials, it is one of many critical care bottlenecks needing adequate provisioning from limited healthcare resources. After observing many deaths in Italy, many experts and government officials are concerned about the number of available ventilators. Individual ones are needed by patients who become critically ill. Equally important resources include medical specialists and staff members that can supervise the patients in their beds and operate their ventilators properly.

Fundamentally, this paper strives to help healthcare managers understand the optimal number of supplies needed for all these lifesaving resources at the peak of the ultimate curve for their aggregate demand. This paper is for two types of audiences. The first audience, familiar with non-stationary queueing theory, can appreciate the closed form formulas and explicit results for time lags that occur within our models. It is critical that our model is non-stationary since stationary models would not produce such insights. The second audience understands how these models describe the dynamics of infectious disease outbreaks and provide insights about important quantities such as delays and time lags between peak demand or patient arrival rates and peak supply or maximum customer usage. Our goal is to have both audiences benefit from this work.

1.1 Main contributions of this paper

We summarize the contributions of this work as follows:

• Derive an expression for the mean number of infected individuals as modelled by an infinite server queue with a non-stationary arrival rate.

• Derive the exact time of the peak load and the value of the peak load for this model.

• Use our peak load calculations to determine the optimal degree of “flattening” needed to make the peak load infection value less than a pre-specified level.

• Study the impact of the duration of the virus on the peak load dynamics of hospital resources and how it affects our flattening policies.

Finally, characterize the nonlinear relationships between

• Flattening the curve.

• The time lag between the peak rate of the newly admitted patients.

• The peak demand for hospital resources.

We also show examples of these time lags being observed in real data.

2 Demand for hospital resources: The infinite server queue

In standard epidemiologic models (SEIR), the rate of spread of infection is represented by the reproductive number [3] (2.4 for the alpha strain of COVID), which itself is related to the product of the number of exposures to an infected individual and the probability of infection, if exposed. This leads to an exponential growth model [4]. In a queueing context, this sets the stage for a Poisson arrival distribution. Interventions aimed at reducing the reproductive number focus on either reducing exposures (e.g. lockdowns and closure of public spaces and
transportation) or reducing the probability of infection (e.g. social distancing and masking). If the reproductive number is lowered to a value of one, then infections are endemic. Similarly, if this number is below one, then infections decrease. This “flattens” the shape of the curve for the number of infections over time.

Over an infinite time period, the total number of people infected is represented by the area under the curve(s). Examples of things that might change the total number of infections would be reducing the susceptible pool (e.g. vaccination) or reducing the duration of infection, returning people more quickly to the susceptible pool. We do not assume that the “flattened” curve in our model changes the total number infected. By assuming a fixed total number, we are considering how capacity constraints can be addressed solely through changes in the rate of arrival. In this sense, it is a conservative approach which ignores the possibility that capacity relief will occur through any of these other possibilities.

We model the number of infectious patients as an $M(t)/G/1$ queueing model. The $M(t)$ represents a time-varying arrival process, $G$ represents a general service distribution and $1$ represents the number of servers. There are two important reasons to begin with this queueing model. First, it is tractable since the distribution is known in closed form for any non-stationary arrival rate. Second, it models the idealized goal of immediate service where no one waits. From the perspective of the COVID-19 epidemic, this means that anyone who wants a bed or a ventilator gets them immediately. These infinite server queues model the maximum resource demand for more realistic systems that may have a finite number of servers with impatient customers. Hence, they represent the dynamics of an unconstrained manager having access to an infinite amount of resources. Finally, they have a history of being used to staff finite server systems, (see for example [5–8]) and serve as the first step towards understanding more complex healthcare systems.

2.1 The $M(t)/G/∞$ queue

This section states the closed form formulas for the $M(t)/G/∞$ queueing model, using the results of [9, 10]. The paper of [9] uses the properties of the Poisson arrival process and Poisson random measures to show that the number in the queue $Q(t)$, has a Poisson distribution with time varying mean $q(t)$. As observed in [9, 10], $q(t)$ has the following explicit solutions

$$q(t) = E[Q(t)] = E \left[ \int_{t-s}^{t} \lambda(u) du \right] = \lambda(t-S) \cdot E[S],$$

where $\lambda(u)$ is the time varying arrival rate at time $u \leq t$. The random variables $S$ and $S_e$ represent, respectively, the customer service time and its long term residual service time. The distribution for $S_e$ is called the stationary excess distribution and is defined to be

$$P\{S_e \leq t\} \equiv \frac{1}{E[S]} \int_{0}^{t} P\{S > u\} du, \quad \text{for all } t \geq 0.$$

Observe that assuming the distribution for $S$ is exponential is equivalent to stating that $S$ and $S_e$ have the same distribution. Moreover, saying that $S$ is a constant is equivalent to $S_e$ being uniformly distributed over a fixed finite positive interval starting at zero. Also note that [11] shows that the first moment of $S_e$ can be expressed in terms of the first and second moments of $S$ since

$$E[S_e] = \frac{E[S^2]}{2E[S]} = \frac{\text{Var}[S]}{2E[S]} + \frac{E[S]}{2} \geq \frac{E[S]}{2}.$$

This lower bound for $E[S_e]$ is tight since it can be attained whenever $S$ is a constant.
A special case for time-varying arrival rates is when \( \lambda(t) = 0 \) for all \( t \leq 0 \) and \( Q(0) = 0 \). The latter condition is a degenerate case for a Poisson distribution and we obtain the formulas

\[
q(t) = E[Q(t)] = E \left[ \int_{(t-S_e)}^{t} \lambda(u) \, du \right] = E[\lambda(t-S_e): S_e \leq t] \cdot E[S].
\]

Hence we only need to make use of this arrival rate function when it is defined on the positive real line.

Finally, as a point of emphasis, note that the word “curve” has two distinct usages:

- Arrival rate functions or service demand.
- Mean queue lengths, queueing loads, or resource demand.

When using infinite server queues to model hospital operations, customer demand maps over to the daily infection rate and resource demand maps over to the daily number of hospital patients. Moreover, there is a causal relationship between the two curves, i.e. you need to flatten the customer demand in order to flatten the resource demand.

Also note that our definition of flattening the curve differs from its typical media usage. In our setting, the average number of people infected over the duration of the epidemic equals a fixed finite number. The resulting peak infection rate is lower in the flattened case and more spread out. We want to understand the capacity usage under different arrival rate shapes. This is a worst-case analysis since a more realistic scenario would have curve flattening leading to an overall lowering of infection numbers.

### 2.2 The \( M_{\phi}(t)/G/\infty \) queue with a scaled Gaussian density arrival rate

In this section, we describe the dynamics of the \( M_{\phi}(t)/G/\infty \) queue whose arrival rate is driven by the Gaussian density function as shown below:

\[
\lambda_{\phi}(t) \equiv \frac{\lambda^*}{\sigma} \cdot \phi \left( \frac{t - \tau}{\sigma} \right) \quad \text{where} \quad \phi(x) \equiv \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2} \quad \text{and} \quad \max_{-\infty < x < \infty} \phi(x) = \frac{1}{\sqrt{2\pi}}. \quad (1)
\]

Given the properties of a Gaussian distribution, \( \tau \) is the location in time for the peak arrival rate, and \( \sigma \) is the standard deviation from this peak time. Thus, \( \lambda^* \), equals the total mean number of infected individuals over the lifetime of the epidemic and the peak arrival rate is

\[
\max_{-\in\infty \leq t \leq \infty} \lambda_{\phi}(t) = \frac{\lambda^*}{\sigma \sqrt{2\pi}}.
\]

The modelling interpretations of \( \lambda^* \), \( \tau \), and \( \sigma \) are summarized by the following integral relationships:

\[
\lambda^* = \int_{-\infty}^{\infty} \lambda_{\phi}(t) \, dt, \quad \tau = \frac{1}{\lambda^*} \int_{-\infty}^{\infty} t \cdot \lambda_{\phi}(t) \, dt, \quad \text{and} \quad \sigma^2 = \frac{1}{\lambda^*} \int_{-\infty}^{\infty} (t - \tau)^2 \cdot \lambda_{\phi}(t) \, dt.
\]

In Fig 1, we plot two scaled Gaussian density arrival rate functions. The blue curve is a scaled Gaussian(\( \tau = 10, \sigma = 2 \)) density arrival function and the red curve (flattened curve) corresponds to the other scaled Gaussian(\( \tau = 20, \sigma = 4 \)) density. Increasing the standard deviation by a factor of 2 reduces the peak value of the arrival rate by the same factor. This follows from the peak value of Gaussian being inversely proportional to the standard deviation \( \sigma \). Our increased understanding of this scaled Gaussian density arrival rate function and using it to model hospital patient arrivals, coupled with insights from these non-stationary queues, leads
to new insights into the dynamics for the total number of people infected and how long they stay hospitalized.

Theorem 2.1. The queue length of an $M_{\phi}(t)/G/1$ queue has a Poisson distribution with mean $q(t)$, which is the solution to the following ordinary differential equation

$$q'(t) = -\frac{\lambda_{\phi}^*}{\sigma^2} \cdot E\left[\frac{t - \tau - S_x}{\sigma} \cdot \phi\left(\frac{t - \tau - S_x}{\sigma}\right)\right] \cdot E[S]$$

$$= -\frac{q(t)}{\sigma} \cdot E\left[\frac{t - \tau - S_x}{\sigma} \cdot \phi\left(\frac{t - \tau - S_x}{\sigma}\right)\right] \cdot E\left[\phi\left(\frac{t - \tau - S_x}{\sigma}\right)\right]^{-1}.$$ (2)

The solution is given by

$$q(t) = \frac{\lambda_{\phi}^*}{\sigma} \cdot E\left[\phi\left(\frac{t - \tau - S_x}{\sigma}\right)\right] \cdot E[S]$$ (3)

and for any value of $t$, we have that

$$q(t) \leq \max_{-\infty < \tau < \infty} q(t) = \frac{\lambda_{\phi}^*}{\sigma} \cdot E[S] \cdot \frac{1}{\sqrt{2\pi}}.$$
Proof. We actually start with the solution of the queue length given in Eq 3. The solution is easily given by the formula from [9] since we have that

\[
q(t) = E[\lambda(t - S)] \cdot E[S] = \frac{\lambda^*}{\sigma} \cdot E \left[ \phi \left( \frac{t - \tau - S}{\sigma} \right) \right] \cdot E[S].
\] (4)

To find the differential equation, take the derivative of the solution with respect to the time parameter \( t \). The final expression in Eq 4 follows from the derivative of the standard Gaussian distribution yielding the first Hermite polynomial times itself i.e. \( \phi'(x) = -x \phi(x) \). Finally, the tight upper bound on the standard Gaussian density function yields similar bounds for the mean queue length. This completes the proof.

We can compute an expression for the time of occurrence of the peak mean queue length. In this paper, we use the time of the peak mean queue length of our model to represent this time. We know that this time for the peak arrival rate occurs at the mode \( \tau \). Our goal is to understand the time lag effect. We define this from when the peak arrival rate occurs until the time of the peak mean queue length. Theorem 2.2 suggests that the time of the peak mean queue length occurs after the peak arrival time \( \tau \). This well known principle in the queueing literature becomes an important predictor within the dynamics for the COVID-19-like outbreak. These results help to determine exactly from the peak number of infections, when the peak load occurs in healthcare systems. The New York Times article [12] has observed in Fig 2 an empirical time lag effect within COVID-19 data. Moreover, this time lag effect persists throughout the empirical data analysis for citywide infection and ICU admissions rates as well. This is seen in Fig 3.

**Theorem 2.2.** The time of the peak mean queue length is the solution to the following fixed point equation

\[
t^* = \tau + E \left[ S_e \cdot \phi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right] \cdot E \left[ \phi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right]^{-1},
\]

which is equivalent to the equation

\[
t^* = \tau + \frac{\lambda^*}{\sigma} \cdot E[S] \cdot \phi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \cdot E[\phi \left( \frac{t^* - \tau - S_e}{\sigma} \right)]^{-1}.
\]

**Proof.** The proofs are found in the Appendix.

**Theorem 2.3.** The time of occurrence of the peak mean queue length is bounded by the following upper and lower bounds

\[
\tau \leq t^* \leq \tau + \frac{E[S]}{\sqrt{2\pi} \cdot \sigma} \cdot E \left[ \phi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right]^{-1}.
\]

**Proof.** Using the first fixed point equation from Theorem 2.2, observe that the numerator is bounded by the standard Gaussian bound given in Eq 1.

Theorem 2.3 provides upper and lower bounds on the time of the occurrence of the peak mean queue length. We can then use a bisection algorithm to determine the time of peak infections exactly. Once the time of the peak infection rate is known, Theorem 2.1 determines both the mean queue length value at that peak time and the value of the peak infection rate. These results provide estimates on the number of ventilators or beds needed at these times.

Now let \( C \) represent this maximum value for the mean queue length at the time of peak infections. With this maximum mean queue length bound, we now show how to flatten the maximum mean queue length of an \( M_{\phi}(t)/G/\infty \) queue using the parameter \( \sigma \). This is the appropriate parameter to use to control the variability with the total mean number of infected staying constant.
Theorem 2.4.

Let $C$ denote the maximum of the mean queue length of an $M_{\sigma}(t)/G/\infty$. The choice of parameter $\sigma$ that achieves this value for $C$ is given by the formula:

$$\sigma^*(C) \equiv \frac{\lambda_s \cdot E[S]}{C \sqrt{2\pi}}.$$
Proof. It follows from the standard Gaussian density function that
\[
\max_{-\infty < t < \infty} q(t) \leq \frac{\lambda^* \cdot E[S]}{\sigma \sqrt{2\pi}}.
\]

Setting this upper bound for the mean queue length equal to the value \(C\) and solving for the standard deviation gives us \(\sigma^*(C)\).

Since properly initialized \(M(t)/G/\infty\) queues always have a Poisson distribution, the following probability is approximately
\[
P\left\{ Q(t^*) > E(Q(t^*)) + \Phi^{-1}(\varepsilon) \cdot \sqrt{E(Q(t^*))} \right\} = P\left\{ \frac{Q(t^*) - E(Q(t^*))}{\sqrt{E(Q(t^*))}} > \Phi^{-1}(\varepsilon) \right\} \approx \varepsilon,
\]
courtesy of the central limit theorem. We call this lower bound for \(Q(t^*)\) an approximate \(\varepsilon\)-threshold capacity. It is an upper bound for \(Q(t^*)\) with an approximate probability of \(1 - \varepsilon\) at least. Moreover \(\varepsilon \leq 1/2\) implies that \(\Phi^{-1}(\varepsilon) \geq 0\), hence
\[
P\left\{ Q(t^*) > \frac{\lambda^* \cdot E[S]}{C \sqrt{2\pi}} + \Phi^{-1}(\varepsilon) \cdot \sqrt{\frac{\lambda^* \cdot E[S]}{C \sqrt{2\pi}}} \right\}
\leq P\left\{ Q(t^*) > E(Q(t^*)) + \Phi^{-1}(\varepsilon) \cdot \sqrt{E(Q(t^*))} \right\}.
\]

Combining these results gives us the following theorem.

Fig 3. The time lag effect via the New York times data.

https://doi.org/10.1371/journal.pone.0286501.g003
Theorem 2.5. Given some value \( C \), an approximate \( \epsilon \)-threshold flattening for an \( M_\nu(t)/G/\infty \) queue can be achieved by setting the parameter \( \sigma \) equal to

\[
\sigma^*(C, \epsilon) = \frac{\lambda^*_\nu \cdot E[S]}{\sqrt{2\pi}} \cdot \left( \frac{\Phi^{-1}(\epsilon) + \sqrt{\Phi^{-1}(\epsilon)^2 + 4C}}{2C} \right)^2.
\]

Proof. Define \( x \) to equal the following constant

\[
x = \sqrt{\frac{\sigma^*(C, \epsilon) \cdot \sqrt{2\pi}}{\lambda^*_\nu \cdot E[S]}} \cdot \frac{\Phi^{-1}(\epsilon) + \sqrt{\Phi^{-1}(\epsilon)^2 + 4C}}{2C}.
\]

Using the ratio of the last term, we can infer from the quadratic formula that \( x \) is a strictly positive real root for the polynomial equation

\[
C \cdot x^2 - \Phi^{-1}(\epsilon) \cdot x - 1 = 0 \quad \text{or equivalently} \quad C = \frac{1}{x^2} + \Phi^{-1}(\epsilon) \cdot \frac{1}{x}.
\]

The latter equation is the \( \epsilon \)-threshold formula

\[
C = \frac{\lambda^*_\nu \cdot E[S]}{\sigma^*(C, \epsilon) \cdot \sqrt{2\pi}} + \Phi^{-1}(\epsilon) \cdot \sqrt{\frac{\lambda^*_\nu \cdot E[S]}{\sigma^*(C, \epsilon) \cdot \sqrt{2\pi}}}
\]

and completes the proof.

Note that when \( \epsilon = 1/2 \), we have \( \Phi^{-1}(1/2) = 0 \). We can then recover the mean result given in Theorem 2.4. Moreover, as \( \epsilon \) gets smaller, we need more flattening to achieve the desired capacity results. Finally, note that this formula also provides an approximate prediction interval for the actual stochastic queue length based on controlling the parameter \( \sigma \).

2.2.1 Deterministic service distribution. In this section, we specify the service distribution to be a constant. This assumption simplifies many of the our equations to yield simpler insights. Our first result shows the mean queue length can be described by an ordinary differential equation (ode) and the solution is simply an integral of the arrival rate from the current time minus the constant service rate until the current time. Finally, we also show that the time lag between the peak arrival rate and peak load is equal to half of the service time.

Corollary 2.6 The mean queue length of an \( M_\nu(t)/D/\infty \) is the solution to the ordinary differential equation

\[
\dot{q}(t) = \lambda_\nu(t) - \lambda_\nu(t - \Delta) = \frac{\lambda^*_\nu}{\sigma} \cdot \varphi\left(\frac{t - \tau}{\sigma}\right) - \frac{\lambda^*_\nu}{\sigma} \cdot \varphi\left(\frac{t - \tau - \Delta}{\sigma}\right)
\]

and the solution is given by

\[
q(t) = \int_{t-\Delta}^t \frac{\lambda^*_\nu}{\sigma} \cdot \varphi\left(\frac{s - \tau}{\sigma}\right) ds = \lambda^*_\nu \cdot \left( \Phi\left(\frac{t - \tau}{\sigma}\right) - \Phi\left(\frac{t - \tau - \Delta}{\sigma}\right) \right)
\]

Finally, the time of the peak is equal to

\[
t^* = \tau + \frac{\Delta}{2}.
\]

Proof. This follows from standard analysis and the symmetry of the Gaussian distribution. It is also important to note that the time derivative of the mean queue length is equal to zero at the midpoint \( \Delta/2 \).
Corollary 2.7. The mean queue length for a mixture of $M\varphi(t)/D/\infty$ queueing models has the following formula

$$q(t) = \sum_{j=1}^{N} \rho_j \tilde{\lambda}_j \cdot \left( \Phi \left( \frac{t - \tau_j}{\sigma_j} \right) - \Phi \left( \frac{t - \tau_j - \Delta}{\sigma_j} \right) \right)$$ (7)

where the $\rho_j$s are the weights of each mixture distribution.

Proof. This follows immediately from the thinning property for non-homogeneous Poisson process.

2.2.2 Exponential service distribution. When the service time distribution is exponential, the mean queue length $q(t)$ at time $t$, solves the autonomous ordinary differential equation

$$-0.4p(t) \dot{q}(t) = \lambda\varphi(t) - \mu \cdot q(t).$$

This linear ordinary differential equation has a closed form solution given by

$$q(t) = q(0) \cdot e^{-\mu t} + \int_{0}^{t} \lambda\varphi(s) \cdot e^{-\mu(t-s)} ds.$$ 

Corollary 2.8. The $M\varphi(t)/M/\infty$ queue has the following closed form expression for its transient mean queue length

$$q(t) = \tilde{\lambda} \cdot e^{-\mu t} \cdot e^{\mu^2 \sigma^2 / 2} \cdot \Phi \left( \frac{t - \tau - S}{\sigma} \right).$$

Proof.

$$q(t) = E[\lambda\varphi(t - \tau - S)] \cdot E[S]$$

$$= \frac{\lambda\varphi}{\sigma} \cdot \mu \sigma \cdot e^{-\mu(t-\tau)} \cdot \sigma \cdot \Phi \left( \frac{t - \tau - \mu \sigma}{\sigma} \right) \cdot \frac{1}{\mu}$$

$$= \frac{\lambda\varphi}{\sigma} \cdot e^{-\mu(t-\tau)} \cdot \sigma^2 \cdot \Phi \left( \frac{t - \tau - \mu \sigma}{\sigma} \right).$$

Corollary 2.9. The mean queue length for a mixture of $M\varphi(t)/M/\infty$ queueing models has the following formula

$$q(t) = \sum_{j=1}^{N} \rho_j \tilde{\lambda}_j \cdot e^{-\mu(t-\tau)} \cdot e^{\mu^2 \sigma^2 j / 2} \cdot \Phi \left( \frac{t - \tau_j - \mu \sigma_j}{\sigma_j} \right)$$ (8)

where the $\rho_j$s are the weights of each mixture distribution.

Theorem 2.10. When the arrival rate function is a scaled Gaussian density and the service time has an exponential distribution, then the lag $\ell$ for the time between the peak arrival rate and the location of the peak mean queue length is given by

$$\ell \equiv t^* - \tau = \sigma \cdot (\mu \sigma + \psi(\mu \sigma)),$$

where

$$\psi^{-1}(x) = \frac{\varphi(x)}{\Phi(x)}.$$
Thus, for the exponential distribution, we have an explicit formula for the time lag in terms of the $\psi(\cdot)$ function. This function is related to conditional expectations of Gaussian random variables and is related to the inverse Mills ratio i.e.

$$
\psi^{-1}(x) = \frac{\varphi(x)}{\Phi(x)} = \frac{E[\mathcal{G} | x + \mathcal{G} \geq 0]}{P(x + \mathcal{G} \geq 0)} = \frac{E[\mathcal{G} \cdot \{ x + \mathcal{G} \geq 0 \}]}{P(x + \mathcal{G} \geq 0)},
$$

where $\mathcal{G}$ is a standard (mean zero, variance one) Gaussian random variable. In the context of operations research, the $\psi(\cdot)$ function has been analyzed in resource sharing settings in [13, 14], as well as, in the context of using risk measures for staffing multi-server queueing systems in [15].

In Fig 4, we plot the same two scaled Gaussian density arrival rate functions given in Fig 1. However, to the right of each arrival curve, we also plot the subsequent mean queue length assuming an exponential service distribution. The parameters for each curve and the service distributions are given in Table 1. Table 1 and Fig 4 yield many observations.

First, note that both arrival rate curves, scaled by the mean service time $1/\mu$, intersect with their mean queue length curves at the time for their peak mean queue length. Since we know the time derivative of the mean queue length is equal to zero this time, then

$$
\dot{q}(t) = \lambda \psi(t) - \mu \cdot q(t) = 0
$$

implies at the time $t^*$ of the occurrence of the peak mean queue length, we must have

$$
\lambda \psi(t^*) = \mu \cdot q(t^*).
$$

(9)

Fig 4. Scaled Gaussian density arrival curves and their mean queue lengths ($E[S] = 1, \lambda^* = 100$ and exponential service distributions).

https://doi.org/10.1371/journal.pone.0286501.g004
Second, for the case of $E[S] = 1$, the peak arrival rate is larger than the peak mean queue length. For the flattened arrival curve, its peak mean queue length value is closer to the peak arrival rate than in the non-flattened case. This implies that flattening the curve and reducing the peak arrival rate by one half, does not necessarily reduce the peak mean queue length by one half. For this example, the peak mean queue length is 91% of the peak arrival rate in the non-flattened curve, while it is 97% of the flattened curve.

In Figs 5 and 6, we plot our Scaled Gaussian density arrival curves with their respective mean queue lengths. In Fig 5, we set the mean service time to be $E[S] = 2$ and in Fig 6, we set the mean service time to be $E[S] = 10$. On the left-hand sides of Figs 5 and 6, we plot the actual arrival rate and the actual mean queue length. However, on the right-hand sides of Figs 5 and 6, we plot the actual mean queue length, but with an arrival rate that is rescaled by the mean service time $E[S]$, i.e. $\lambda(t) \cdot E[S]$. We observe that for the plots on the left, the arrival rate does not intersect with the mean queue length at the peak. However, for the rescaled plots on the right, observe that the arrival rate intersects with the mean queue length curve at the peak of the latter. This verifies Eq 9 again. Finally, we observe in Table 1 that by reducing the arrival rate peak by one half reduces the mean queue length peak by one half when $E[S] = 1$. However, we observe in Tables 2 and 3 that as we increase $E[S]$, we see a smaller reduction in the peak mean queue length. Certainly we expect the infectious period to be longer than one day for COVID-19 so our theoretical results imply that we need to do more flattening of the curve when patients spend more time infected.

**Remark.** Poisson thinning, when the probabilities for thinning are not a function of time, does not affect the time of the peak arrival rate. The time of the peak load is unaffected whenever the thinning mechanism is also independent of the random service time. In general, we can thin the number of infected individuals appropriately to consider the number of people who enter a hospital to make use of an ICU or not and in either case recover or not.

A constant arrival rate over the entire real line generates an $M/G/\infty$ queue in steady state. The superposition property of Poisson processes and infinite server queues adds a constant to the arrival rate, which does not change these times either. Hence scaling the total number of infected people down, due to flattening the curve, has no effect on the location of the peak mean queue length.

### 2.3 The $M_\gamma(t)/G/\infty$ queue with a scaled gamma density arrival rate

We now describe $M_\gamma(t)/G/\infty$ queueing dynamics when the arrival rate $\lambda_\gamma(t)$ is a scaled gamma distribution function of time $t$ or

$$
\lambda_\gamma(t) \equiv \frac{\lambda^\circ \cdot \beta^\circ}{\Gamma(x)} \cdot t^{x-1} \cdot e^{-\beta^\circ t}, \text{ for all } t \geq 0, \text{ where } \Gamma(x) = \int_0^\infty e^{-t^x} dt.
$$

Table 1. Comparison between regular and flattened curves peak values and times ($\lambda^\circ = 100$) and exponential service distributions.

| Curve Type                  | $\tau$ | $\sigma$ | $E[S]$ | Peak Time | Peak Value |
|-----------------------------|--------|----------|--------|-----------|------------|
| Arrival Curve               | 10     | 2        | 1      | 10        | 19.95      |
| Flattened Arrival Curve     | 20     | 4        | 1      | 20        | 9.97       |
| Mean Queue Length           | 10     | 2        | 1      | 10.86     | 18.20      |
| Flattened Mean Queue Length | 20     | 4        | 1      | 20.95     | 9.70       |

https://doi.org/10.1371/journal.pone.0286501.t001
Fig 5. Scaled Gaussian density arrival curves and their mean queue lengths ($E[S] = 2, \lambda^* = 100$). Arrival Rate (Left). Rescaled Arrival Rate (Right).

https://doi.org/10.1371/journal.pone.0286501.g005
Here $\alpha$ is a shape parameter for the gamma distribution, $\beta$ is the rate parameter for the gamma distribution, and $\lambda^*_g$ is again the total mean number of infected individuals over the lifetime of the epidemic. Unlike the standard Gaussian case, the gamma density is only guaranteed to be bounded when $\alpha \geq 1$. Since it is unrealistic to assume infinite infection rates, we

![Diagram](https://doi.org/10.1371/journal.pone.0286501.g006)
Table 2. Comparison between regular and flattened curves peak values and times ($\lambda^*_\alpha = 100$).

| Curve Type                  | $\tau$ | $\sigma$ | $E[S]$ | Peak Time | Peak Value |
|-----------------------------|--------|----------|--------|-----------|------------|
| Arrival Curve               | 10     | 2        | 2      | 10        | 19.95      |
| Flattened Arrival Curve     | 20     | 4        | 2      | 20        | 9.97       |
| Mean Queue Length           | 10     | 2        | 2      | 11.40     | 31.29      |
| Flattened Mean Queue Length | 20     | 4        | 2      | 21.71     | 18.20      |

https://doi.org/10.1371/journal.pone.0286501.t002

Table 3. Comparison between regular and flattened curves peak values and times ($\lambda^*_\alpha = 100$).

| Curve Type                  | $\tau$ | $\sigma$ | $E[S]$ | Peak Time | Peak Value |
|-----------------------------|--------|----------|--------|-----------|------------|
| Arrival Curve               | 10     | 2        | 2      | 10        | 19.95      |
| Flattened Arrival Curve     | 20     | 4        | 2      | 20        | 9.97       |
| Mean Queue Length           | 10     | 2        | 2      | 11.40     | 31.29      |
| Flattened Mean Queue Length | 20     | 4        | 2      | 21.71     | 18.20      |

https://doi.org/10.1371/journal.pone.0286501.t003

assume this constraint on $\alpha$ for the remainder of this paper. This assumption, implies

$$\max_{-\infty < \phi < \infty} \lambda^*_\phi(t) \leq \lambda^* \left(\frac{\alpha - 1/\beta}{\beta} \right) = \frac{\lambda^*_\alpha \cdot \beta}{\Gamma(\alpha)} \cdot \left(\frac{\alpha - 1}{e}\right)^{\alpha - 1} \leq \frac{\lambda^*_\alpha \cdot \beta}{\sqrt{2\pi\alpha}},$$  \hspace{1cm} (10)

where the final inequality follows from Stirling’s approximation. The properties of gamma densities give us the following integral relationships:

$$\lambda^*_\phi = \int_0^\infty t \lambda(t) dt, \quad \frac{\alpha}{\beta} = \frac{1}{\lambda^*_\phi} \int_0^\infty t \cdot \lambda(t) dt, \quad \text{and} \quad \frac{\alpha}{\beta^2} = \frac{1}{\lambda^*_\phi} \int_0^\infty \left(t - \frac{\alpha}{\beta^2}\right)^2 \lambda(t) dt.$$

The standard deviation for the spread of the arrival rate distribution is now $\sqrt{\alpha}/\beta$. Hence, we have an upper bound for the peak arrival rate that is inversely proportional to the spread of the arrival rate function. Note that this bound has the same form as for the scaled Gaussian arrival rate case.

One reason for using the gamma density is that it is not symmetric about its mean like the Gaussian one. We have data examples where the arrival rate function is asymmetric and scales to a distribution with a heavy tail. One reason for this is that the number of cases may go towards zero quite slowly i.e. until a vaccine is available in a particular country. These data-driven insights encourage us to explore the scaled gamma density arrival rate function.

In Fig 7, we plot two scaled gamma density arrival rate functions. The blue curve uses a Gamma($\alpha = 5, \beta = .5$) density and the red curve (flattened curve) uses a Gamma($\alpha = 10, \beta = .5$) density. By increasing the variance by a factor of 2, we have reduced the peak value of the arrival rate by a factor of two. We now leverage this analysis of the scaled gamma density arrival rate function, for a Poisson patient arrival model, to obtain new insights from non-stationary queues on the dynamics of the total number infected and their total infection time.

**Theorem 2.11.** The $M(t)/G/\infty$ queueing model has a Poisson distribution with mean $q(t)$. Moreover, this mean solves the ordinary differential equation

$$\dot{q}(t) = \frac{\lambda^* \cdot \beta^2}{\Gamma(\alpha)} \cdot E[(\alpha - 1 - \beta \cdot (t - S_t)) \cdot (t - S_t)^{\alpha - 2} \cdot e^{-\beta(t-S_t)}; S_t \leq t] \cdot E[S]$$

$$= \frac{E[(\alpha - 1 - \beta \cdot (t - S_t)) \cdot (t - S_t)^{\alpha - 2} \cdot e^{-\beta(t-S_t)}; S_t \leq t]}{E[(t - S_t)^{\alpha - 1} \cdot e^{-\beta(t-S_t)}; S_t \leq t]} \cdot q(t)$$  \hspace{1cm} (11)
and the solution is given by

\[ q(t) = \frac{\lambda^* \beta^\alpha}{\Gamma(\alpha)} \cdot \text{E}[(t - S_e)^{\alpha-1} e^{-\beta(t-S_e)}; S_e \leq t] \cdot \text{E}[S]. \]  

Moreover, for all \( \alpha \geq 1 \),

\[ \max_{-\infty < t < \infty} q(t) \leq \frac{\lambda^* \beta}{\Gamma(\alpha)} \left( \frac{\alpha - 1}{e} \right)^{\alpha-1} \cdot \text{E}[S] \leq \frac{\lambda^* \beta}{\sqrt{2\alpha \pi}} \cdot \text{E}[S]. \]

**Proof.** We actually start with the solution. The solution is easily given by the formula from [9]. To find the differential equation, take the derivative of the solution with respect to the time parameter \( t \). The resulting bound for the gamma density function yields the corresponding bounds for the mean queue length.

**Theorem 2.12.** The time of the peak mean queue length solves the following fixed point equation

\[ t^* = \frac{\alpha - 1}{\beta} + \frac{\text{E}[S_e \cdot (t^* - S_e)^{\alpha-2} e^{-\beta(t^*-S_e)}; S_e \leq t^*]}{\text{E}[(t^* - S_e)^{\alpha-2} e^{-\beta(t^*-S_e)}; S_e \leq t^*]}. \]

**Proof.** To find the time of the peak mean queue length, set the time derivative of the mean queue length to zero, i.e.

\[ \dot{q}(t^*) = 0. \]
Using the differential equation given in Eq 11, we then have

\[
\frac{\lambda^*_e \cdot \beta^e}{\Gamma(x)} \cdot E\left[ (x - 1 - \beta \cdot (t^* - S_e)) \cdot (t^* - S_e)^{x-2} \cdot e^{\beta \cdot (t^* - S_e)}; S_e \leq t^* \right] \cdot E[S] = 0.
\]

Isolating \( t^* \) by itself on the left hand side gives us the formula.

Note how this result decomposes into a sum of two distinct parts. The first summand is the mode for the arrival rate function, which for the gamma distribution occurs at \( (\alpha - 1)/\beta \). The second summand is the positive shift from this mode. This also implies that the time of peak mean queue length lags after the time of the peak arrival rate just as in the Gaussian case.

Fig 8 plots our scaled gamma density arrival curves with their respective mean queue lengths. On the left of Fig 8, we set the mean service time to be \( E[S] = 1 \) and on the right of Fig 8, we set the mean service time to be \( E[S] = 10 \). Finally, we observe in Table 1 that by reducing the arrival rate peak by one half reduces the mean queue length peak by one half when \( E[S] = 1 \). However, we observe in Tables 4 and 5 that as we increase \( E[S] \), there is a smaller reduction in the peak mean queue length. We expect the infectious period to be longer than one day for COVID-19, so our theoretical results imply that we may need to do more flattening of the curve as patients spend more time infected.

3 Understanding COVID-19 through data

In this section, we use some of the infection and death COVID-19 data made available through the Johns Hopkins University website [16]. Fig 9 plots the number of daily confirmed infections in Italy the and United States. Note how the confirmed cases start earlier in Italy than in the United States. Also observe that a local infection rate maximum occurs on the date of April 16th, 2020. For each of the plots, we also overlay properly scaled Gaussian and gamma distribution density functions to get a sense what distribution parameters best fit the data. For Italy, the best scaled Gaussian fit is \( \lambda^* = 3679154 \) times a Gaussian(\( \tau = 177, \sigma = 19 \)) density and the best scaled gamma fit is \( \lambda^* = 57824445 \) times a Gamma(\( \alpha = 36.4, \beta = .2 \)) density. For the United States, the best scaled Gaussian fit is \( \lambda^* = 1377772 \) times a Gaussian(\( \tau = 60, \sigma = 83 \)) density and the best scaled gamma fit is \( \lambda^* = 286662 \) times a Gamma(\( \alpha = 13, \beta = .2 \)) density.

Finally, note that we used the “fitdist” function in Matlab for fitting the data to the scaled Gaussian and gamma distributions respectively.

Fig 10 plots the daily arrivals and deaths from January 22nd, 2020 to October 4th, 2020 normalized by their maximum value (arrival = black dots, deaths = red dots). For Italy and the United States, the maximum confirmed infections was (6557) and (77,362) respectively. Moreover, for Italy, and the United States, the maximum confirmed deaths was (919), and (2609) respectively. Observe that only for the United States does the time for the largest number of deaths lag the same for the largest number of confirmed infections. This is the time lag effect observed in our queueing models. This lag also inspires our construction of a new way to view it.

Fig 11 plots the cumulative distribution functions for the total number of arrivals and deaths in Italy and the United States. This plot matches the quantiles of the arrivals to the quantiles to that of the deaths. As can be observed from the left-hand plot of Fig 11, the deaths in Italy lagged behind the infections up until April 2020. However the post-April deaths start to slow down and the total death quantiles start leveling off to 1 and the total number of deaths started to lag behind the total number of infections. For the United States, by contrast, the total number of infections overtook the total number of deaths but total death quantiles have not leveled off. Both countries have been able to slow down their death rates but Italy has seemingly done a better job in limiting the number of deaths.
Fig 8. Scaled Gamma density arrival curves and their mean queue lengths $E[S] = 2$ (Left) and $E[S] = 10$ (Right).

https://doi.org/10.1371/journal.pone.0286501.g008
4 Key takeaways and insights

The analysis of the queueing model provides both qualitative insights and prescriptive guidance on how to flatten the demand for hospital resources.

4.1 Flattening the curve

First, we show that the peak number of hospitalized patients is inversely related to flattening the rate of arriving patients (Theorem 2.1). This insight provides a mechanism to control the peak number of hospitalized people through the standard deviation parameter $\sigma$. As a result, flattening should be inversely proportional to the amount of hospital capacity to avoid overcrowding. However, the effectiveness of social distancing on flattening the curve, i.e. increasing $\sigma$, remains an open question. Additionally, we find that flattening the curve has a smaller impact on the peak number of patients when the service times are long. This means the longer patients stay in the hospital, the more society must flatten the curve in order to achieve the same capacity level. In the three examples with an exponential service distribution and a scaled Gaussian density arrival rate, we flattened the peak arrival rate by 50%, however, we see that the peak mean queue length was reduced by 47%, 42%, 23% respectively for $E[S]$ equaling 1, 2, and 10. For the scaled gamma density arrival rate setting, a 33% reduction in the peak arrival rate obtains a 31% and 16% reduction respectively for $E[S] = 1$ and $E[S] = 10$.

4.2 Time lag between the peak infection rate and the peak number of patients

A key takeaway from our queueing analysis is that there is a lag between the time of the peak arrival rate of infected patients does not coincide with the time of the peak of number of hospitalized people (or deaths). We characterize this time lag exactly for the simplified queue model (Theorem 2.2). First, we show that the peak number of infected lags behind the peak number of newly infected individuals. Surprisingly, there is a nonlinear relationship between flattening the curve and the time lag between the peak rate of newly admitted patients and the peak demand for hospital resources (Theorem 2.2). Moreover, we show in Theorem 6.3 that the
Fig 9. Arrival scatter plots.

https://doi.org/10.1371/journal.pone.0286501.g009
Fig 10. Arrival and death scatter plots.

https://doi.org/10.1371/journal.pone.0286501.g010
time lag $\ell$ is bounded by the mean service time for an exponential distribution i.e.

$$\frac{1}{\mu} - \frac{1}{\mu^2 \sigma^2 + \mu} = \frac{1}{\mu} \frac{(\mu \sigma)^2}{(\mu \sigma)^2 + 1} \leq \frac{1}{\mu} = E[S].$$

This analysis also implies that as we flatten the curve i.e. increase $\sigma$, the time lag converges to the mean service time $1/\mu$. 

Fig 11. CDF of arrival and death numbers.
https://doi.org/10.1371/journal.pone.0286501.g011
5 Conclusion

In this paper, we present a simple infinite server queueing model for the number of infected people with Covid-19. This analysis can be used to predict how and when many people will become infected, need a hospital bed, or need a ventilator in the near future. We show how the dynamics of the number infected explicitly depends on the arrival rate function. We explore the case of a symmetric rate function by using a scaled Gaussian density function. Our asymmetric example is a scaled gamma density function. We explicitly calculate for both cases the times of their peak mean queue lengths and the peak mean queue lengths, along with their dependence on the duration of infection.

To address more of the health operations issues for the larger pandemic event, we outline below some of the remaining important questions that still need to be studied. By leveraging operations research tools and techniques, hospitals have prided themselves on the efficiency that comes with just-in-time supply management and minimizing empty beds, thus saving on hospital costs. However as a result, they are often not equipped for an epidemic surge such as this one. Despite our efforts to understand the impact of COVID-19 on the healthcare system, there is plenty of additional research that needs to be done in order to understand these emerging complexities.

One area of importance is to understand the effects of surges in arrivals into the healthcare system. Some recent research has studied queueing models with self-exciting point processes that model random surges in demand, see for example [15, 17–21]. These types of models would generalize our model in this paper but they are more complicated and would disguise the easy, back of the envelope insights from simpler models.

Future research should address the impact of testing and time lags in testing for the coronavirus. It has been shown through empirical analyses that testing delays can have a huge impact on the reported numbers of positive cases and health outcomes, see for example [22]. Our density plots in Fig 11 show that testing is an important indicator for predicting deaths to follow.

Finally, we are observing in some cities that it is important to use demographic information for understanding the impact of COVID-19 on particular communities or regions of the country. In the city of Chicago, despite African-Americans making up roughly 30% of the Chicago population, African-Americans make up 52% of cases of COVID-19 and 68% of COVID-19 deaths, see for example [23]. Also in NYC, it has been reported that COVID-19 has a disproportionately negative impact on low-income communities [24]. Recent work by [25] combines queueing theory with demographic information to get a deeper understanding of blood donation dynamics. This suggests that similar demographic analyses are also needed to understand the full impact of COVID-19 on marginalized populations.

6 Appendix

6.1 The $M_\varphi(t)/G/\infty$ queue

6.1.1 Proof of Theorem 2.2. $\text{Proof.}$ In order to find the peak mean queue length we need to set the time derivative of the mean queue length to zero i.e.

$$\dot{q}(t^*) = 0.$$
Now using the differential equation given in Eq 2, we have

\[ \hat{q}(t^*) = E \left[ (t^* - \tau - S_e) \cdot \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right] = 0. \]

This implies that

\[ (t^* - \tau) \cdot E \left[ \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right] = E \left[ S_e \cdot \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right]. \]

Now isolate \( t^* \) by itself to get

\[ t^* = \tau + E \left[ S_e \cdot \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right]^{-1} \]

\[ = \tau + \lambda_s^* \cdot E[S] \cdot E \left[ S_e \cdot \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right]. \]

This completes the proof.

6.1.2 Discrete service distribution. Corollary 6.1. The mean behavior for the \( M_\phi(t)/D_n/\infty \) queueing model is the solution to the ordinary differential equation

\[ \hat{q}(t) = \sum_{i=1}^{n} p_i \cdot (\lambda_\phi(t) - \lambda_\phi(t - \Delta)) \]

\[ = \sum_{i=1}^{n} p_i \cdot \left( \frac{\lambda^*_s}{\sigma} \cdot \varphi \left( \frac{t - \tau}{\sigma} \right) - \frac{\lambda_\phi^*}{\sigma} \cdot \varphi \left( \frac{t - \tau - \Delta}{\sigma} \right) \right) \quad (13) \]

and the solution is given by

\[ q(t) = \sum_{i=1}^{n} p_i \cdot \left( \int_{t - \Delta}^{t} \frac{\lambda_\phi}{\sigma} \cdot \varphi \left( \frac{s - \tau}{\sigma} \right) ds \right) \]

\[ = \sum_{i=1}^{n} p_i \cdot \lambda_\phi \cdot \left( \Phi \left( \frac{t - \tau}{\sigma} \right) - \Phi \left( \frac{t - \Delta - \tau}{\sigma} \right) \right). \quad (14) \]

Proof. The proof follows the thinning of Poisson processes and the expression given in Eq 6.

6.1.3 Exponential service distribution. Theorem 2.10

Proof. First we observe that from Theorem 2.2 that \( t^* \) solves the following fixed point equation

\[ t^* = \tau + E \left[ S_e \cdot \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{t^* - \tau - S_e}{\sigma} \right) \right]^{-1}. \]

If we define the peak lag time to be the parameter \( \ell \), we then have \( \ell = t^* - \tau \) and the following fixed point equation

\[ \ell = E \left[ S_e \cdot \varphi \left( \frac{\ell - S_e}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{\ell - S_e}{\sigma} \right) \right]^{-1}. \]
Thus, if we define $X$ to be a standard (mean 1) exponential random variable, then we have

$$
\ell = E \left[ S_e \cdot \varphi \left( \frac{\ell - S_e}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{\ell - S_e}{\sigma} \right) \right]^{-1}
$$

$$
= E \left[ \frac{X}{\mu} \cdot \varphi \left( \frac{\ell - X/\mu}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{\ell - X/\mu}{\sigma} \right) \right]^{-1}
$$

$$
= \mu \sigma \cdot e^{-\mu \ell} \cdot e^{-\ell^2/2} \cdot \left( (\ell - \mu \sigma^2) \cdot \Phi \left( \frac{\ell - \mu \sigma}{\sigma} \right) + \sigma \cdot \varphi \left( \frac{\ell - \mu \sigma}{\sigma} \right) \right)
$$

$$
= \ell - \mu \sigma^2 + \sigma \cdot \varphi \left( \frac{\ell}{\sigma} - \mu \sigma \right) \cdot \Phi \left( \frac{\ell}{\sigma} - \mu \sigma \right)^{-1}
$$

Subtracting the time lag $\ell$ from both sides and using the $\psi$ function gives us

$$
\mu \sigma = \sigma \cdot \varphi \left( \frac{\ell}{\sigma} - \mu \sigma \right) \cdot \Phi \left( \frac{\ell}{\sigma} - \mu \sigma \right)^{-1}
$$

if and only if $\psi(\mu \sigma) = \frac{\ell}{\sigma} - \mu \sigma$.

This gives us an explicit solution for $\ell$ which is

$$
\ell = \sigma \cdot (\mu \sigma + \psi(\mu \sigma)).
$$

Now it remains to prove two identities from the next lemma.

**Lemma 6.2** We have the following two identities for standard exponential random variables:

$$
E \left[ \varphi \left( \frac{\ell - X/\mu}{\sigma} \right) \right] = \mu \sigma \cdot e^{-\mu \ell} \cdot e^{-\ell^2/2} \cdot \Phi \left( \frac{\ell}{\sigma} - \mu \sigma \right)
$$

and

$$
E \left[ \frac{X}{\mu} \cdot \varphi \left( \frac{\ell - X/\mu}{\sigma} \right) \right]
$$

$$
= \mu \sigma \cdot e^{-\mu \ell} \cdot e^{-\ell^2/2} \cdot \left( (\ell - \mu \sigma^2) \cdot \Phi \left( \frac{\ell}{\sigma} - \mu \sigma \right) + \sigma \cdot \varphi \left( \frac{\ell}{\sigma} - \mu \sigma \right) \right).
$$

**Proof.** For the first equality, define $\Theta$ to be a standard (mean 0 and variance 1) Gaussian random variable. Exploiting the version of Stein’s lemma as given in [21, 26] for standard Gaussian random variables $\Theta$ along with its analogue for exponential random variables, we have for all suitable functions $f$

$$
E[\Theta \cdot f(\Theta)] = E[f'(\Theta)] \quad \text{and} \quad E[f(\Theta)] = f(0) + E[f'(\Theta)].
$$
Assuming that $\mathcal{G}$ and $\mathcal{X}$ are independent random variables and applying Stein’s formula to generalized functions, gives us

$$
E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] = E \left[ E \left( \mathcal{G} \cdot \{ \mathcal{G} > \frac{\ell - \mathcal{X}/\mu}{\sigma} | \mathcal{X} \} \right) \right] \\
= E \left[ E \left( \mathcal{G} \cdot \{ \mathcal{G} > \frac{\ell - \mathcal{X}/\mu}{\sigma} | \mathcal{G} \} \right) \right] \\
= E \left[ \mathcal{G} \cdot P \left\{ \mathcal{G} > \frac{\ell - \mathcal{X}/\mu}{\sigma} \right\} \right].
$$

Since $\mathcal{X}$ is a standard exponential, we use Stein’s formula again and obtain

$$
E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] = E \left[ \mathcal{G} \cdot P \left\{ \mathcal{X} > \mu \cdot \left( \frac{\ell - \sigma \mathcal{G}}{\sigma} \right) \right\} \right] \\
= E \left[ \mathcal{G} \cdot \left( e^{-\mu \left( \frac{\ell - \sigma \mathcal{G}}{\sigma} \right)} \cdot \left\{ \mathcal{G} \leq \frac{\ell}{\sigma} \right\} + \left\{ \mathcal{G} \leq \frac{\ell}{\sigma} \right\} \right) \right] \\
= E \left[ \mathcal{G} \cdot \left( e^{-\mu \left( \frac{\ell - \sigma \mathcal{G}}{\sigma} \right)} - 1 \cdot \left\{ \mathcal{G} \leq \frac{\ell}{\sigma} \right\} \right) \right] \\
= E \left[ \mu \sigma \cdot e^{-\mu \left( \frac{\ell - \sigma \mathcal{G}}{\sigma} \right)} \cdot \left\{ \mathcal{G} \leq \frac{\ell}{\sigma} \right\} \right].
$$

Finally, Girsanov’s formula gives us

$$
E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] = \mu \sigma \cdot e^{-\mu \ell} \cdot E \left[ e^{\alpha \mathcal{G}} \cdot P \left\{ \mathcal{G} + \mu \sigma \leq \frac{\ell}{\sigma} \right\} \right] \\
= \mu \sigma \cdot e^{-\mu \ell} \cdot e^{\sigma^2/2} \cdot \Phi \left( \frac{\ell}{\sigma} - \mu \sigma \right).
$$

Since

$$
\frac{d}{d\ell} E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] = \frac{1}{\sigma} \cdot E \left[ \varphi' \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] \\
= \frac{1}{\sigma} \cdot \left( -\frac{\ell}{\sigma} \cdot E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] + \frac{1}{\sigma} \cdot E \left[ \frac{\mathcal{X}}{\mu} \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] \right),
$$

we now have

$$
\frac{d}{d\ell} \log E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] = \frac{d}{d\ell} E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right]^{-1} \\
= -\frac{\ell}{\sigma^2} + \frac{1}{\sigma^2} \cdot E \left[ \frac{\mathcal{X}}{\mu} \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{\ell - \mathcal{X}/\mu}{\sigma} \right) \right]^{-1}.
$$
Taking the logarithmic derivative for the right hand side of the first identity and then multiplying both sides by \( \sigma^2 \) gives us
\[
-m\sigma^2 + \sigma \cdot \varphi \left( \frac{\ell}{\sigma} - \mu \sigma \right) \cdot \Phi \left( \frac{\ell}{\sigma} - \mu \sigma \right)^{-1} = -\ell + E \left[ \frac{X}{\mu} \cdot \varphi \left( \frac{\ell - X}{\sigma} \right) \right] \cdot E \left[ \varphi \left( \frac{\ell - X}{\sigma} \right) \right]^{-1}.
\]
We then have
\[
E \left[ \frac{X}{\mu} \cdot \varphi \left( \frac{\ell - X}{\sigma} \right) \right] = E \left[ \varphi \left( \frac{\ell - X}{\mu} \right) \right] \cdot \left( \ell - m\sigma^2 + \sigma \cdot \varphi \left( \frac{\ell}{\sigma} - \mu \sigma \right) \cdot \Phi \left( \frac{\ell}{\sigma} - \mu \sigma \right)^{-1} \right)
\]
and the second identity follows from using the first.

**Theorem 6.3** For the function \( \psi(x) \), we have the following bound.
\[
-\frac{1}{x^3 + x} \leq \psi(x) + x - \frac{1}{x} \leq 0.
\]
Thus, we have in the exponential service setting with a scaled Gaussian arrival rate function that the time lag is bounded by
\[
-\frac{1}{\mu} \cdot \frac{1}{(\mu \sigma)^2 + 1} \leq \ell - \frac{1}{\mu} \leq 0.
\]

**Proof.** The first inequality follows from Theorem 7.2 in [14]. Therefore, the time lag bounds follow by substituting \( x = \mu \).

**6.1.4 Hyper-exponential service distribution.** Another continuous distribution of interest is the hyper-exponential distribution, a special case of a phase type distribution. When holding the mean of the distribution fixed, the Erlang distribution always has a smaller variance than the corresponding one for an exponential distribution. The corresponding hyper-exponential distribution always has a larger variance than both of theirs. Usually the hyper-exponential distribution is determined by two \( n \)-dimensional vectors of parameters \( (p_1, p_2, \ldots, p_n) \) and \( (\mu_1, \mu_2, \ldots, \mu_n) \). The first vector represents the probabilistic weights for an exponential distribution associated with the corresponding rate parameter given by the second vector. The hyper-exponential density is then a convex combination of exponential distributions.

**Corollary 6.4.** The \( M_\psi(t)/H_n/\infty \) queue with a scaled Gaussian arrival rate is given by the following closed form expression for the mean transient queue length
\[
q_{H_n}(t) = \sum_{i=1}^{n} p_i \cdot q_i(t),
\]
where
\[
q_i(t) = \sigma \cdot e^{-\mu_i (t-\tau)} \cdot e^{t^2/2} \cdot \Phi \left( \frac{t - \tau}{\sigma} - \mu_i \cdot \sigma \right)
\]
for all \( i = 1, \ldots, n \).

**Proof.** This follows from the fact that the hyper-exponential is just a thinning of a Poisson process. Thus, the sum is also a Poisson process with the sum of the rates.
**Corollary 6.5.** The $\text{M}_{t}/\text{H}_n/\infty$ queue with arrival rate given by a Gaussian distribution with parameters has the following closed form expression for the time lag

$$
\xi = \frac{\sum_{j=1}^{n} \alpha_j \mu_j \cdot e^{-\mu_j t} \cdot e^{\mu_j^2 \sigma^2 / 2} \cdot \Phi \left( \frac{t - \mu_j \cdot \sigma}{\sigma} \right)}{\sum_{j=1}^{n} \alpha_j \mu_j \cdot e^{-\mu_j t} \cdot e^{\mu_j^2 \sigma^2 / 2} \cdot \left( \frac{t - \mu_j \cdot \sigma}{\sigma} \right) \cdot \Phi \left( \frac{t - \mu_j \cdot \sigma}{\sigma} \right) + \sigma \cdot \Phi \left( \frac{t - \mu_j \cdot \sigma}{\sigma} \right)}
$$

where

$$
\alpha_j = \frac{P_j}{\mu_j \cdot \sum_{i=1}^{n} P_i / \mu_i}.
$$

**Proof.** This follows from the fact that the hyper-exponential is just a thinning of a Poisson process that is the superposition of the sum of the rates. Thus, the sum is also a Poisson process with the sum of the rates. Moreover, the stationary excess distribution of a hyper-exponential is another hyper-exponential distribution with modified rates [11].

**Author Contributions**

**Conceptualization:** Jamol J. Pender, William A. Massey, Robert C. Hampshire.

**Data curation:** Sergio Palomo, Jamol J. Pender, William A. Massey, Robert C. Hampshire.

**Formal analysis:** Sergio Palomo, Jamol J. Pender, William A. Massey, Robert C. Hampshire.

**Investigation:** Jamol J. Pender, William A. Massey, Robert C. Hampshire.

**Methodology:** Jamol J. Pender, William A. Massey, Robert C. Hampshire.

**Supervision:** Jamol J. Pender, William A. Massey.

**Writing – original draft:** Jamol J. Pender, William A. Massey.

**Writing – review & editing:** Jamol J. Pender, William A. Massey, Robert C. Hampshire.

**References**

1. Goebel B. California Substantially Flattened the COVID-19 Curve in March; April 6th, 2020. Available from: https://www.noozhawk.com/article/brian_goebel_california_flattened_covid_19_curve_in_march_20200405 [cited April 6th, 2020].

2. Hilk M, Azad A, Bruer W. Every hotspot has ‘its own curve’: How coronavirus cases are growing around the United States. https://www.cnn.com/2020/03/24/health/coronavirus-state-hotspots/indexhtml. 2020.;

3. Adam D. A guide to R—the pandemic’s misunderstood metric. Nature. 2020; 583(7816):346–349. https://doi.org/10.1038/d41586-020-02009-w PMID: 32620883

4. Kendall DG. Deterministic and stochastic epidemics in closed populations. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability. vol. 4. University of California Press Berkeley; 1956. p. 149–165.

5. Jennings OB, Mandelbaum A, Massey WA, Whitt W. Server staffing to meet time-varying demand. Management Science. 1996; 42(10):1383–1394. https://doi.org/10.1287/mnsc.42.10.1383

6. Feldman Z, Mandelbaum A, Massey WA, Whitt W. Staffing of time-varying queues to achieve time-stable performance. Management Science. 2008; 54(2):324–338. https://doi.org/10.1287/mnsc.1070.0821

7. Liu Y, Whitt W. Stabilizing customer abandonment in many-server queues with time-varying arrivals. Operations research. 2012; 60(8):1551–1564. https://doi.org/10.1287/opre.1120.1104

8. Pender J, Massey WA. Approximating and stabilizing dynamic rate Jackson networks with abandonment. Probability in the Engineering and Informational Sciences. 2017; 31(1):1–42. https://doi.org/10.1017/S0269964816000371

9. Eick SG, Massey WA, Whitt W. The physics of the $\text{M}/\text{G}/\infty$ queue. Operations Research. 1993; 41(4):731–742. https://doi.org/10.1287/opre.41.4.731
10. Eick SG, Massey WA, Whitt W. $M_t/G/\infty$ queues with sinusoidal arrival rates. Management Science. 1993; 39(2):241–252. https://doi.org/10.1287/mnsc.39.2.241

11. Massey WA, Whitt W. A probabilistic generalization of Taylor’s theorem. Statistics & probability letters. 1993; 16(1):51–54. https://doi.org/10.1016/0167-7152(93)90122-Y

12. Leatherby L, Sun A. These Graphs Show How COVID-19 Is Ravaging New York City’s Low-Income Neighborhoods; January 21st, 2022. Available from: https://www.nytimes.com/interactive/2022/01/21/us/covid-hospitalizations.html/

13. Hampshire RC, Massey WA, Mitra D, Wang Q. Provisioning for bandwidth sharing and exchange. In: Telecommunications Network Design and Management. Springer; 2003. p. 207–225.

14. Hampshire RC, Massey WA, Wang Q. Dynamic pricing to control loss systems with quality of service targets. Probability in the Engineering and Informational Sciences. 2009; 23(2):357–383. https://doi.org/10.1017/S0269964809000205

15. Pender J. Risk measures and their application to staffing nonstationary service systems. European Journal of Operational Research. 2016; 254(1):113–126. https://doi.org/10.1016/j.ejor.2016.03.012

16. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases. 2020;

17. Daw A, Pender J. Exact simulation of the queue-hawkes process. In: Proceedings of the 2018 Winter Simulation Conference. IEEE Press; 2018. p. 4234–4235.

18. Daw A, Pender J. An ephemerally self-exciting point process. Advances in Applied Probability. 2022; 54(2):340–403. https://doi.org/10.1017/apr.2021.35

19. Daw A, Pender J. Queues driven by Hawkes processes. Stochastic Systems. 2018; 8(3):192–229. https://doi.org/10.1287/stsy.2018.0014

20. Nyirora J, Pender J. Optimal staffing in nonstationary service centers with constraints. Naval Research Logistics (NRL); 2016; 63(8):615–630. https://doi.org/10.1002/nav.21723

21. Massey WA, Pender J. Dynamic Rate Erlang-A Queues. Queueing Systems. 2018; 89(1-2):127–164. https://doi.org/10.1007/s11134-018-9581-2

22. Ryan-Mosley T. How you could get your covid-19 test results faster; April 6th, 2020. Available from: https://www.technologyreview.com/s/615444/covid-19-test-results-faster-commercial-labs-delays-coronavirus/ [cited April 6th, 2020].

23. Reyes C, Husain N, Gutowski C, Clair SS, Pratt G. Chicago’s coronavirus disparity: Black Chicagoans are dying at nearly six times the rate of white residents, data show. Chicago Tribune. 2020; 7.

24. Wilson C. These Graphs Show How COVID-19 Is Ravaging New York City’s Low-Income Neighborhoods; April 20th, 2020. Available from: https://time.com/5821212/coronavirus-low-income-communities/ [cited April 20, 2020].

25. Kahara M, Pender J. Using simulation to study the impact of racial demographics on blood transfusion allocation policies. In: 2017 Winter Simulation Conference (WSC). IEEE; 2017. p. 2786–2797.

26. Massey WA, Pender J. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. Queueing Systems. 2013; 75(2-4):243–277. https://doi.org/10.1007/s11134-012-9340-8