# Appointment Scheduling with a Quantile Objective

Peijun Sang[a], Mehmet A. Begen[b,*], Jiguo Cao[c]

[a]*Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada*
[b]*Ivey Business School, Western University, London, ON, Canada*
[c]*Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada*

**Abstract**

Appointment scheduling has many applications (e.g., surgery scheduling, airport gate scheduling, container vessel dockings and radiation therapy bookings) and it has a direct and significant operational and economic impact. For example, in healthcare, surgical departments are one of the main drivers of hospital costs and revenue, and appointment scheduling is used to book surgeries. Effective scheduling not only enables patients' timely access to care but also enables more efficient operations. This becomes especially important as healthcare costs and demand are on the rise in many countries. We study appointment scheduling where there are jobs (e.g., patients, container vessels, airplanes) with random processing durations, an expensive processor (e.g., a doctor, dock crane, airport gate) and significant costs for processor idle time, processor overtime, and job waiting. The goal is to determine an appointment schedule that minimizes a measure of total costs as the objective. The appointment scheduling problem has been well studied in the literature with the expected cost objective. Almost all papers in the literature on appointment scheduling use the expected cost criterion, which may not be suitable when risk measures and/or service levels are considered. In this paper, we study this problem with a new objective: minimization of any quantile of the cost distribution, e.g., median, 90th percentile. We obtain theoretical results for some special cases and develop an algorithm for the general case. Our algorithm does not require a specific distribution assumption and can work directly with data samples. We present numerical examples with real data on surgeries. Our results show that allocated schedules based on the quantile objective with identical jobs are different than the ones generated by the expected cost objective and they do not show the well-known dome-shaped pattern but a semi-dome-shaped pattern which first increases (like the dome-shaped pattern) but then its decrease is not monotone (unlike the dome-shaped pattern). To the best of our knowledge, this is the first paper on appointment scheduling problem with the objective of the quantile function minimization.

*Keywords:* appointment scheduling, quantile minimization, surgery scheduling, risk measure, median minimization, service level

## 1. Introduction

We determine planned start times (appointment times) of jobs to be processed on a single resource/processor. The job sequence is given. The resource is expensive and processing durations are random. Due to random durations, for a given appointment schedule, there may be some idle time of the processor, some overtime of the processor, and some wait time of jobs. A good schedule minimizes a measure of idle time, wait time and overtime costs (i.e., a measure of the total cost).

Appointment scheduling has many applications in various industries in which it has a direct and significant operational and economic impact. For example, in healthcare appointment scheduling is used to book surgeries

*Corresponding author. Email: mbegen@ivey.uwo.ca

and surgical departments are one of the main drivers of hospitals' costs and revenues ([1, 2, 3]), e.g., estimates of \$2000/hour ([4]) or even \$100/minute ([5]) for regular operating room time are not uncommon. Any savings on operating rooms usage are substantial. Furthermore, effective scheduling not only enables timely access to care but also helps to achieve this in an economical way, and it becomes especially important as healthcare costs are a significant portion of the GDP of most countries ([2, 6]). Canada is expected to spend \$264 billion in 2019 [7], and demand and costs for healthcare services are on the rise in almost all countries, e.g., the GDP portion of healthcare costs in the US is expected to double by 2050 ([1]). Other healthcare applications include physician appointments ([8]), CT scan bookings ([9]), and radiation treatment appointments ([10]). Besides healthcare, there are important applications of appointment scheduling in other industries such as transportation ([11, 12]), supply chain ([13]), call center staffing ([14]), cloud computing ([15]), and other service industries ([16, 17]).

Appointment scheduling is an active area of research and there have been many papers in the last 60 years ([18, 19, 20, 21]). These papers use various methods to determine the optimal appointment durations that minimize the expected of the total cost ([22, 23, 8]) or determine the number of bookings per a fixed period of time to minimize the expected total cost ([24, 25, 26]). The main techniques used are stochastic programming ([23, 27, 8]), newsvendor approaches ([28]), simulation ([29, 30, 31, 32, 33]), optimization-simulation ([34, 35]), dynamic programming ([9, 36, 37]), and queuing theory ([11, 24, 38]). Most of the papers assume that distributions of processing durations are known and available; however, there are some that study appointment scheduling when distributions are not known and only samples are available, or only some partial or limited information is available ([39, 40, 41, 42]).

The appointment scheduling literature, except a few exceptions, focuses on the objective of minimization of the expected (mean) total cost, e.g., see ([22, 18, 1]) and the references therein. The expected cost can be a good measure for some of the applications, e.g., in a setting where a decision maker is solving the same scheduling problem every day again and again with many processors to minimize the expected cost of overtime, idle time and job waiting time. For some applications, especially in the service industry, the expected cost may not be a good measure for an objective. Furthermore, in some applications, instead of the expected value, a risk measure may be more suitable, e.g., a patient clinic may want to schedule their appointments so that 90th percentile of patient wait time is minimized to have a better service level for their patients. Another reason why the expected cost sometimes may not be a good measure could be processing duration distributions. If durations are highly skewed, expectation may not be a fair measure as it fails to take the asymmetrical shape of the distribution into account. In this case, median or a more general quantile function may be a better candidate of the objective function to be minimized. Furthermore, if the underlying distributions of processing durations are not known and only their samples are available to estimate the optimal appointment scheduling, then the median is more robust compared to the mean, i.e., less sensitive to extreme large or small values of the duration time.

In this paper, we introduce a new measure of the total cost as the objective of the appointment scheduling problem - minimization of a quantile (e.g., median, 95th percentile) of the total cost distribution. For example, the objective can be to minimize the 90th percentile of wait time or to minimize the 75th percentile of total cost. We develop methods to solve the appointment scheduling problem with this new objective.

Appointment scheduling received little attention with an objective function other than the minimization of expected of cost. [43] develop a multi objective model using fuzzy logic to generate operating room (OR) schedules. The authors allocate OR capacities to different specialties, which is known as OR block scheduling or master surgical scheduling ([44, 45, 46, 47]). They do not consider appointment scheduling, and they use fuzzy logic to generate robust schedules. [48] develop an optimization-based variability reduction method for surgery scheduling with an objective function minimizing the CVaR (conditional value at risk) of overtime and idle time costs. CVar is more concerned about the weighted average of extreme losses in the right tail of the total cost. This is an important measure if extremely large cost values are of primary interest when scheduling appointments. The quantile function used in our work, however, can provide a more detailed characterization of the distribution of the total cost, rather than just the right tails. Another method, robust optimization, aims to seek a solution being high quality and effective under any scenario (i.e., finding a robust solution) for problems in which the underlying uncertainty can be represented as deterministically in parameter values. The idea is to somehow remove uncertainty by taking into account the worst possible case so that the obtained solution will be robust for any scenario, e.g., minimize max cost. Robust optimization approaches minimize the worst-case scenario, i.e., the 100th percentile of the cost function and it has become an active and popular area of research over the last 30 years ([49, 50, 51, 52]). The need for robust methods can arise from ambiguity and/or uncertainty and sometimes it can help to convert a stochastic model into a deterministic one, and there are many applications in diverse areas such as engineering ([53]), finance ([54, 55, 56]) and healthcare ([57, 58]). [59] develop a robust optimization model for elective admissions in a hospital to reduce bed shortages for emergency patients. Their model is a capacity allocation model; it can be viewed as an advance scheduling model ([9, 60, 61]) and does not consider appointment scheduling. The studies ([62, 63, 64, 65]) develop robust optimization models to solve a variant of the appointment scheduling problem. Our approach differs from robust optimization approaches in that quantile level is given (i.e., it is a choice of the decision maker and it is not necessarily the highest 100th quantile value) and we are not able to convert our model into a deterministic one. In a sense our approach is less conservative. To the best of our knowledge, minimization of a quantile of the cost is a first in the literature for appointment scheduling. Our objective function is also flexible in its coefficients and can represent any combination of wait time, idle time and overtime as desired.

We first analyze a special case of the general problem (a single job with equal cost coefficients and log normal processing distribution), obtain insights and extend our results for a greater number of jobs, generic coefficients, and any distribution. We obtain theoretical results for some special cases and develop methods for the general case. Our methods do not require a specific distribution assumption and can work directly with data samples. We present numerical examples with real data on surgeries. Our results, for identical jobs, show that optimal appointment times exhibit a linear like shape regardless of the quantile level whereas optimal allocated durations do not show the well-known dome-shaped pattern but a semi-dome-shaped pattern which first increases (like the dome-shaped pattern) but then its decrease is not monotone (unlike the dome-shaped pattern).

The paper is organized as follows. In the next section we present our analysis for the single job case. In Section 3, we analyze the two-job case and we generalize to multiple jobs in Section 3.2. We present our numerical examples and results with real data in Section 4. Section 5 concludes the paper.

| | |
|---|---|
| $A_k$ | the appointment time for the $k$th job |
| $U_i$ | the service time for the $i$th job |
| $C_i$ | the completion time of job $i$ |
| $o_i$ | coefficient of the overage cost for job $i$ |
| $u_i$ | coefficient of the underage cost of job $i$ |
| $Y$ | the total of the overage cost and the underage cost |
| $Q_X(q)$ | the $q$th quantile function of a random variable $X$. |

Table 1: Notation used throught the paper.

## 2. Single Job

Let $N$ represent the number of jobs and $A_i$ be the appointment time for job $i$, $i = 1, 2..., N$. (Here $N = 2$, since we have two jobs with $A_1 = 0$ and only one job to schedule, job 2.) We assume that $A_1 = 0$. Let $C_i$ be the completion time of job $i$. Let $U_i$ be the service time for job $i$, then $C_1 = U_1$. For now, we assume that $U_i \sim$ log-normal$(\mu, \sigma^2)$ for $i = 1...N$. We use this assumption for demonstration purposes and relax it later in the paper and can work directly with data samples (i.e., without any distribution information). The motivation for the log-normal distribution assumption is that surgery durations generally are log-normally distributed [66]. Then we can represent the total cost, $Y_1$ as

$$Y_1 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+, \tag{1}$$

where $(a)^+ = \max(0, a)$ and $o_i$ and $u_i$ are coefficients of the overage cost and underage cost for job $i$, $i = 1, 2..., N$, respectively. Let $Q_{Y_1}(A_2|q)$ represent the $q$th quantile function of the total cost $Y_1$, which is a function of $A_2$. Our objective is to find $\hat{A}_2$ that minimizes $Q_{Y_1}(A_2|q)$ for a given probability level $q$. We first analyze this model when $o_1 = u_1$ in the next section.

### *2.1. Equal Cost Coefficient Case*

We consider the case of equal cost coefficients, i.e., $o_1 = u_1 = \alpha$. Then $Y_1 = \alpha|C_1 - A_2|$. We aim to find the value $\hat{A}_2$, which minimizes the quantile of $Y_1$. To find the quantile of $Y_1$, investigation of its distribution is indispensable, as shown in Figure 1.

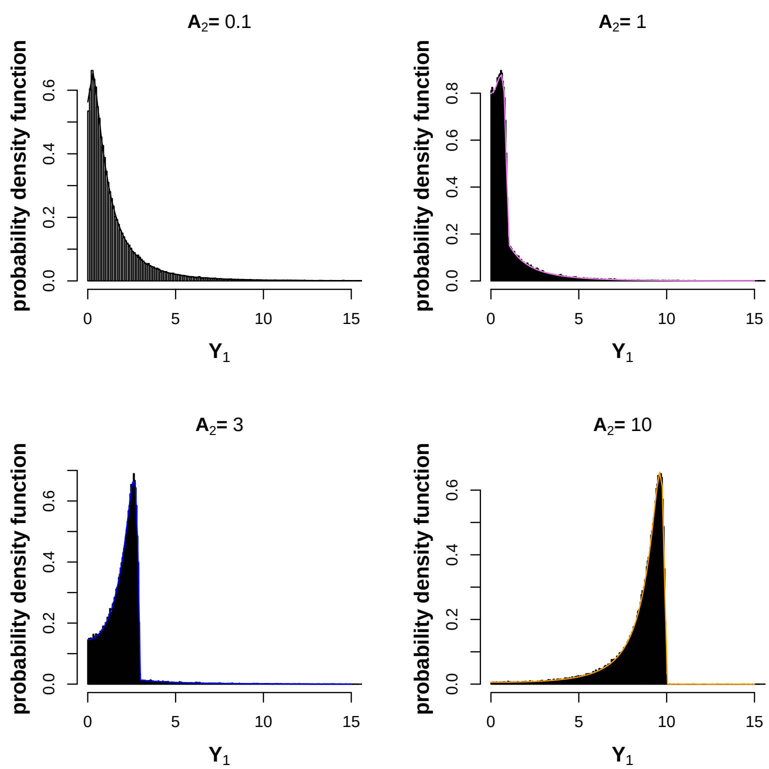



Figure 1: Probability density function of the total cost for different values of $A_2$. For all four panels, we set the parameters as $\alpha = 1,\ \mu = 0,\ \sigma = 1$.

Figure 1 displays the distribution (density) of $Y_1$ with different values of $A_2$ after setting $\mu = 0,\ \sigma = 1$ and $\alpha = 1$. The plots imply that what really affects of the shape of the distribution (left skewed, right skewed, heavy tailed) is the relative magnitudes of $A_2$ and $\mu$. More specifically, the distribution of $Y_1$ switches from right skewed to left skewed as $A_2$ increases. Asymmetry of the distribution of $Y_1$ suggests that the median might be quite different from the mean of the total cost for given values of parameters. Results for other distributions such as a lognormal with mean 20 and standard deviation 5 and an exponential distribution with mean 20 are displayed in Figure 7 in the appendix. The same conclusions hold for these distributions as well. Additionally, we also include the probability density of function of $Y_1$ for unequal overage and underage cost coefficients; please refer to Figure 8 in the appendix. In terms of the distribution of $Y_1$, the same conclusion holds for these extra distributions considered in Figure 8 .

We now present our result to find the optimal appointment time $\hat{A}_2$ that minimizes $Q_{Y_1}(A_2|q)$ for a given probability level $q$.

**Proposition 1.** $\hat{A}_2(q) = \frac{Q_{U_1}(\hat{x})+Q_{U_1}(\hat{x}+q)}{2}$, where $\hat{x} = \arg\min_x\{Q_{U_1}(x+q) - Q_{U_1}(x)\}$.

*Proof.* We have

$$\begin{aligned}
&P(Y_1 \le Q_{Y_1}(A_2|q)) \\
&= P(\alpha|C_1 - A_2| \le Q_{Y_1}(A_2|q)) \\
&= P\left(A_2 - \frac{Q_{Y_1}(A_2|q)}{\alpha} \le C_1 \le A_2 + \frac{Q_{Y_1}(A_2|q)}{\alpha}\right) \\
&= F_{C_1}\left(A_2 + \frac{Q_{Y_1}(A_2|q)}{\alpha}\right) - F_{C_1}\left(A_2 - \frac{Q_{Y_1}(A_2|q)}{\alpha}\right) \\
&= q,
\end{aligned}$$

where $F_{C_1}$ denotes the cumulative distribution function of $C_1$. Let $F_{C_1}\left(A_2 - \frac{Q_{Y_1}(A_2|q)}{\alpha}\right) = x$, then $F_{C_1}\left(A_2 + \frac{Q_{Y_1}(A_2|q)}{\alpha}\right) = x + q$, where $x \in (0, 1-q)$. Since $U_1 = C_1$,

$$\begin{aligned}
F_{U_1}\left(A_2 - \frac{Q_{Y_1}(A_2|q)}{\alpha}\right) &= x \\
F_{U_1}\left(A_2 + \frac{Q_{Y_1}(A_2|q)}{\alpha}\right) &= x + q,
\end{aligned} \tag{2}$$

Equation (2) implies that $A_2 - \frac{Q_{Y_1}(A_2|q)}{\alpha}$ and $A_2 + \frac{Q_{Y_1}(A_2|q)}{\alpha}$ are the $x$th and $(x+q)$th quantiles of $U_1$, respectively. Let $Q_{U_1}(x)$ be the $x$th quantile of $U_1$. Then

$$\begin{aligned}
Q_{U_1}(x) &= A_2 - \frac{Q_{Y_1}(A_2|q)}{\alpha}; \\
Q_{U_1}(x+q) &= A_2 + \frac{Q_{Y_1}(A_2|q)}{\alpha}; \\
A_2 &= \frac{Q_{U_1}(x) + Q_{U_1}(x+q)}{2}.
\end{aligned}$$

Note that the length of the interval $[Q_{U_1}(x), Q_{U_1}(x+q)]$ is $2/\alpha \cdot Q_{Y_1}(A_2|q)$, and the middle point of the interval $[Q_{U_1}(x), Q_{U_1}(x+q)]$ is $A_2$. In order to find $\hat{A}_2$ that minimizes $Q_{Y_1}(A_2|q)$ for a given probability level $q$, it suffices to find a $\hat{x}$, which has the shortest interval $[Q_{U_1}(\hat{x}), Q_{U_1}(\hat{x}+q)]$, since the middle point of the shortest interval $[Q_{U_1}(\hat{x}), Q_{U_1}(\hat{x}+q)]$ is the $\hat{A}_2$ that minimizes $Q_{Y_1}(A_2|q)$ for a given probability level $q$. Then the optimal appointment time is

$$\hat{A}_2(q) = \frac{Q_{U_1}(\hat{x}) + Q_{U_1}(\hat{x}+q)}{2} \tag{3}$$

where $\hat{x} = \arg\min_x\{Q_{U_1}(x+q) - Q_{U_1}(x)\}$. □

**Remark 1.** *Unfortunately there is no closed form of the quantile function $Q_{U_1}(x)$ or the difference in the two quantile functions $Q_{U_1}(x+q) - Q_{U_1}(x)$ for a general log-normal distribution. To see this, suppose $U_1$ follows a log-normal distribution $(\mu, \sigma^2)$, we have*

$$\begin{aligned} x &= P(U_1 \leq Q_{U_1}(x)) \\ &= P(\log(U_1) \leq \log(Q_{U_1}(x)) \\ &= P\left(\frac{\log(U_1) - \mu}{\sigma} \leq \frac{\log(Q_{U_1}(x)) - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{\log(Q_{U_1}(x)) - \mu}{\sigma}\right), \end{aligned}$$

*where $Z$ follows the standard normal distribution $N(0,1)$. In other words, $\frac{\log(Q_{U_1}(x))-\mu}{\sigma} = \Phi^{-1}(x)$, where $\Phi^{-1}(x)$ satisfies*

$$\int_{-\infty}^{\Phi^{-1}(x)} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = x.$$

*There is no closed form for $\Phi^{-1}(x)$ hence there is no closed form for $Q_{U_1}(x)$.*

**Remark 2.** *For log-normal distribution, there is an easy way to solve the $Q_{U_1}(x+q) - Q_{U_1}(x)$ minimization problem. According to Theorem in [67], the minimizer $\hat{x}$ satisfies that*

$$Q_{U_1}(\hat{x}) = \exp(\mu + \sigma A), \quad Q_{U_1}(\hat{x}+q) = \exp(\mu + \sigma B),$$

*where $A$ and $B$ are the unique solutions of the two equations*

$$\begin{aligned} \Phi(B) - \Phi(A) &= q \\ A + B &= -2\sigma, \end{aligned}$$

*and $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$. To solve these two equations, we can use numerical methods such as the bisection method since there is no closed form solution.*

If we take another distribution for appointment durations, for instance, exponential distribution with a rate parameter $\lambda$, we can show that the difference in quantiles is convex.

**Proposition 2.** *$Q_{U_1}(x+q) - Q_{U_1}(x)$ is convex if appointment duration is exponentially distributed.*

*Proof.* For an exponential distribution with a rate parameter $\lambda$ we have

$$\int_0^{Q_{U_1}(x)} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda Q_{U_1}(x)} = x,$$

which leads to $Q_{U_1}(x) = -\log(1-x)/\lambda$. Therefore,

$$Q_{U_1}(x+q) - Q_{U_1}(x) = -\log(1-x-q)/\lambda + \log(1-x)/\lambda.$$

Its second derivative is $1/(1-x-q)^2 - 1/(1-x)^2 > 0$ for $x \in (0, 1-q)$. $\square$

Figure 2 displays $Q_{U_1}(x+q) - Q_{U_1}(x)$ changing with $x$ when the value of $q$ varies while setting $\mu = 4$ and $\sigma = 1$. All functions $Q_{U_1}(x+q) - Q_{U_1}(x)$ have a minimum. As the probability level $q$ (or the corresponding quantile) increases, the minimizer of $Q_{U_1}(x+q) - Q_{U_1}(x)$ decreases accordingly. This result is verified to hold not only for some asymmetric distributions such as the log-normal distribution, but also for symmetric distributions such as the normal distribution. This is demonstrated in Figure 9 in the appendix, where $U_1$ is assumed to follow a normal distribution with mean 20 and standard deviation 5.

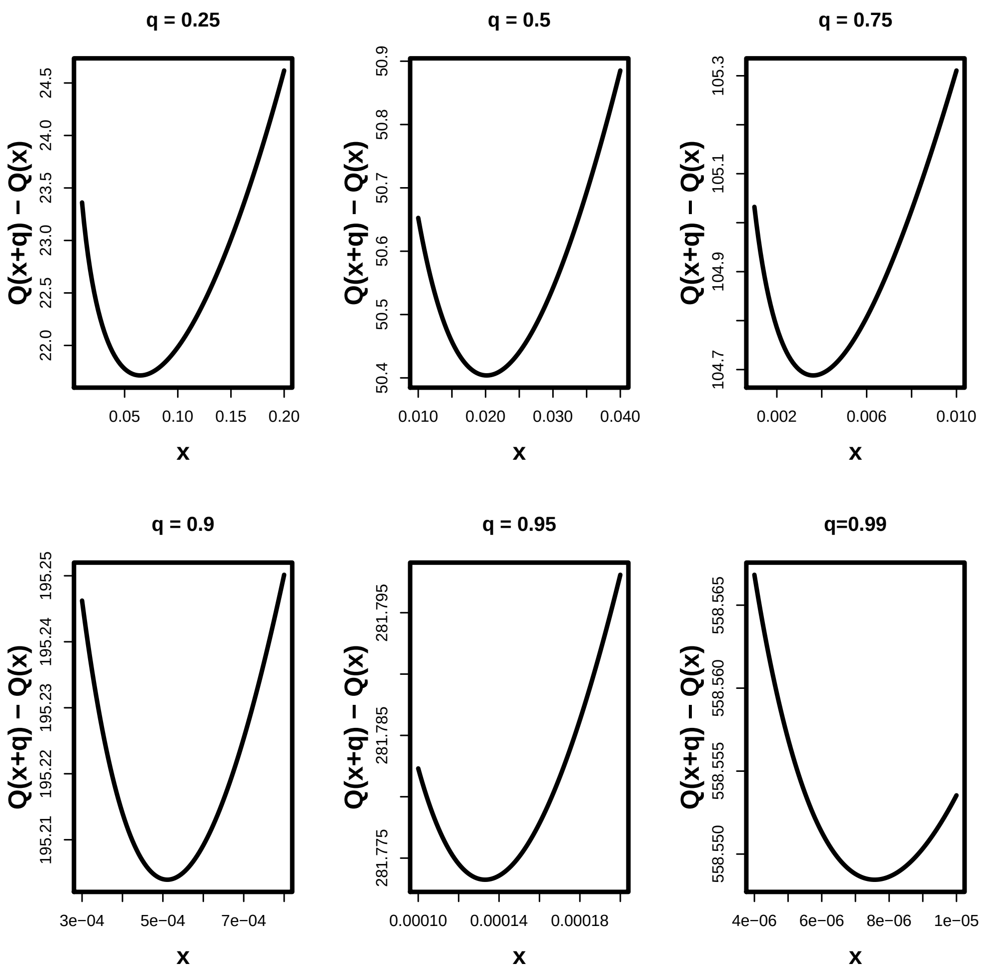



Figure 2: $Q_{U_1}(x+q) - Q_{U_1}(x)$ changes with $x$ when the value of $q$ varies. The service duration is assumed to follow log-normal(4, 1).

We develop a Newton-type optimization algorithm to find $\hat{x}$ and implement all of our methods in R ([68]). We provide a trace of the optimization algorithm over iterations in Figure 10 when $q = 0.5$ and $q = 0.75$ in the appendix. The implementation details of this algorithm are available in Algorithm 1 in the appendix.

| | $q = 0.50$ | $q = 0.75$ | $q = 0.95$ |
|---|---|---|---|
| $x_0 = 10^{-2}$ | 0.59 | 1.03 | 2.61 |
| $x_0 = 10^{-3}$ | 0.59 | 1.03 | 2.61 |
| $x_0 = 10^{-4}$ | 0.59 | 1.03 | 2.61 |

Table 2: The optimal appointment time $\hat{A}_2$ in the single-job scheduling with various starting values ($x_0$) and various probability levels ($q$). We set the constant parameters as $\alpha = 2$, $\mu = 0$, and $\sigma = 1$.

Our algorithm is stable in terms of the choice of the starting value $x_0$ for evaluating the function $Q_{U_1}(x +$

$q) - Q_{U_1}(x)$. For different starting values, the corresponding optimal appointment time remains the same as long as the cost coefficients $\alpha$, $\mu$ and $\sigma$ are fixed. As shown in Table 2, where we set the constant parameters $\alpha = 2$, $\mu = 0$, and $\sigma = 1$, the corresponding optimal appointment time does not change with different starting values. In addition, the algorithm is able to converge in a small number of steps, as shown in Figure 10 of the appendix.

We show how the optimal appointment time changes with respect to $q$ and $\mu$ in Figure 3. In the left panel of Figure 3, the optimal appointment time, $\hat{A}_2(q)$ is plotted against probability (or quantiles) $q$ under different values of $\mu$ while setting $\alpha = 2$ and $\sigma = 1$. For all three cases, as the probability (or quantile) $q$ increases, the corresponding optimal appointment time $\hat{A}_2(q)$ increases accordingly. In addition, the shape of the function $\hat{A}_2(q)$ is insensitive to the choice of $\mu$. However, the magnitude of $\mu$ can affect the magnitude of the optimal appointment time. To make this effect more explicit, the right panel of Figure 3 displays the optimal appointment time $\hat{A}_2(0.5)$, $\hat{A}_2(0.75)$ and $\hat{A}_2(0.95)$ as a function of $\mu$ when minimizing the 0.5th, 0.75th and 0.95th quantiles of the total cost. As $\mu$ increases, the corresponding optimal appointment time increases as well. In both cases (increasing $q$ or $\mu$), the increases in the optimal appointment time are easy to interpret. Even though $\hat{x}$, which minimizes $Q_{U_1}(x+q)-Q_{U_1}(x)$, decreases when increasing $q$, the overall effect on $Q_{U_1}(\hat{x}+q)+Q_{U_1}(\hat{x})$ is increasing. In other words, the decrease in the quantile $Q_{U_1}(\hat{x})$ cannot compensate for the increase in the quantile $Q_{U_1}(\hat{x}+q)$ as $q$ increases. As for the other case, when $\mu$ is increased, more probability is put on larger values of the service duration. Keeping $q = 1/2$, the corresponding quantile $Q_{U_1}(\hat{x} + q)$ will increase even though the minimizer $\hat{x}$ of $Q_{U_1}(x + q) - Q_{U_1}(x)$ decreases. This explains why the corresponding optimal appointment time, a multiple of $Q_{U_1}(\hat{x}+q)+Q_{U_1}(\hat{x})$, increases as $\mu$ increases. We consider another scenario in which the duration time follows an exponential distribution with $\mu$ being the mean. The corresponding optimal appointment times under different $\mu$ or $q$ are displayed in Figure 11 in the appendix. Furthermore, a similar pattern in terms of the relationship between optimal appointment time and $\mu$ or $q$ can be found in that figure.

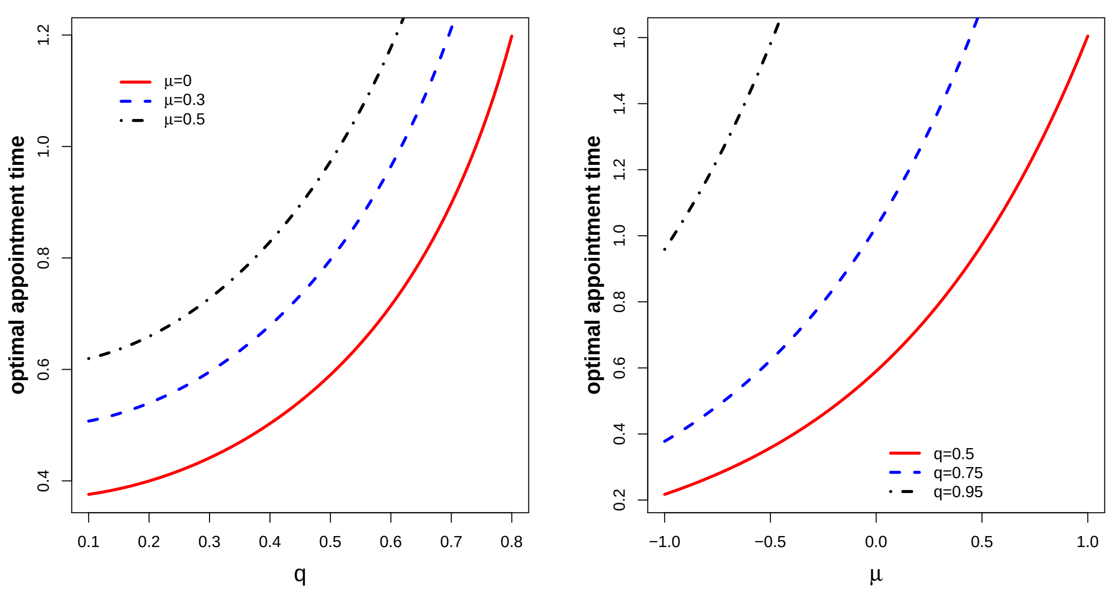



Figure 3: The optimal appointment time under different $\mu$ and $q$ values. For both left and right panels, we set the constant parameters as $\alpha = 2$ and $\sigma = 1$.

In contrast to the relationship between the optimal appointment time and $\mu$, the relationship between $\hat{A}_2(q)$ and $\sigma$ is less obvious. The left panel of Figure 4 displays $\hat{A}_2(q)$ as a function of $q$ under different values of $\sigma$ while setting $\mu = 0$ and $\alpha = 2$. In all three scenarios, when $\sigma = 0.5, 0.7, 1$, the corresponding optimal appointment time $\hat{A}_2(q)$ increases as the probability level $q$ increases. As $q$ increases, even though the minimizer $\hat{x}$ will decrease (and thus, $Q_{U_1}(\hat{x})$ will decrease accordingly), the decrease of this quantile cannot compensate for the increase of the other quantile $Q_{U_1}(\hat{x} + q)$. This explains why the resulting optimal appointment time increases as $q$ increases. However, when $q$ is fixed, the relationship between the optimal appointment time and $\sigma$ is more complex. Comparing these three lines, we see that when $q$ is relatively small - say less than 0.6 - decreasing $\sigma$ may result in the increase of the optimal appointment time. In other words, if our interest focuses on the lower to moderate quantile of total cost, decreasing variation of the (log) service duration would lead to a larger optimal appointment time. For large quantiles - say larger than the 0.8th quantile of the total cost - decreasing the variation of the (log) service duration will yield a smaller optimal appointment time. These cases are considerably more ambiguous for $q$ between 0.6 and 0.8. The right panel of Figure 4 displays the optimal appointment time $\hat{A}_2(q)$ changing with $\sigma$ when $q = 0.5,\ 0.75,\ 0.95$, which confirms the above conclusion. Moreover, $\hat{A}_2(0.5)$ keeps decreasing at about a constant rate at relatively small values of $\sigma$. As $\sigma$ becomes increasingly larger, the decrease of $\hat{A}_2(0.5)$ slows down accordingly. Besides this experiment, we considered another one where $U_1$ follows a lognormal distribution with $\mu = 3$ and various $\sigma$'s; the corresponding optimal appointment times with respect to different combinations of $q$ and $\sigma$ are displayed in Figure 12 in the appendix. This figure shows a similar pattern in comparison with Figure 4.

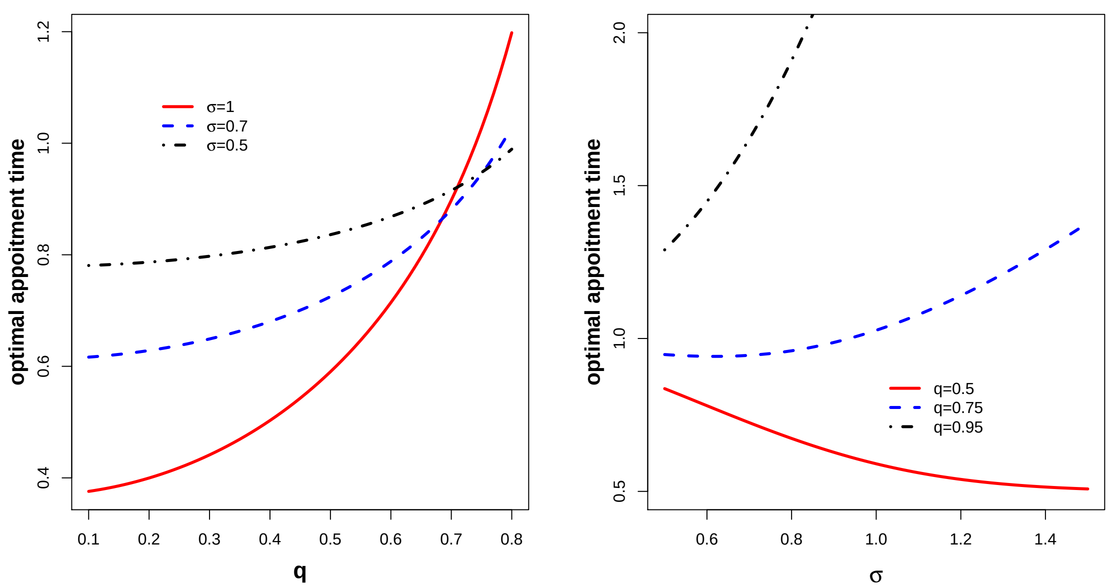



Figure 4: The optimal appointment time under different $\sigma$ and $q$ values. For both left and right panels, we set the constant parameters as $\alpha = 2$ and $\mu = 0$.

*2.2. Sampling Approach*

In some cases, theoretical quantile functions of the total cost may be intractable or difficult to evaluate. For example, the underlying distribution of the service duration may be unknown and only some independent observations of the service duration may be available. Another case is when the cost coefficients are not equal in (1) (i.e., $o_1 \neq u_1$). (See Section 2.3 for the unequal cost coefficients.) In these cases, employing the sample quantile of the total cost to replace the population quantile might be a shortcut to yield the optimal appointment time, which is close to the true optimal appointment time (based on minimizing population quantile) if the sample size is large enough. Note that this sampling method can be applied for any quantile and any distribution.

Let $U_{11}, \ldots, U_{1n}$ be a random sample of the service duration $U_1$. For a given $A_2$, the corresponding total cost, $Y_{11}, \ldots, Y_{1n}$, can be obtained from (1), if the cost coefficients are known. Then we are able to evaluate the sample $q$th quantile of $Y_{1i}, \ldots, Y_{ni}$. This procedure is, in spirit, similar to using a sample mean as a surrogate of expectation of a distribution when the exact distribution is not known. From this perspective, we regard the $q$th quantile of the total cost as a (univariate) function of $A_2$. Now our goal is to find the optimal appointment time $\hat{A}_2$, which can minimize this $q$th quantile of the total cost. Many approaches are available to find a minimum of a univariate function. We use a Newton-type algorithm (Algorithm 1 in the appendix).

Table 3 summarizes the comparison of the sample-based optimal appointment time and the real optimal appointment time. The sample-based optimal appointment time is based on optimization for the 0.2th quantile, the median and the 0.8th quantile of the total cost, assuming an empirical distribution for the service duration. The real optimal appointment time is acquired from the algorithm introduced in Section 2.1, while assuming the underlying distribution of the service duration is the log-normal distribution. We set the constant parameters as $\alpha = 2$, $\mu = 0$ and $\sigma = 1$. If the sample size is sufficiently large, the sample-based optimal appointment time is close to the real optimal appointment time.

| | $n = 100$ | $n = 1000$ | $n = 10000$ | *Real* |
|---|---|---|---|---|
| $q = 0.2$ | 0.50 | 0.45 | 0.39 | 0.40 |
| $q = 0.5$ | 0.66 | 0.67 | 0.60 | 0.59 |
| $q = 0.8$ | 1.25 | 1.25 | 1.23 | 1.20 |

Table 3: Comparison of the sample-based optimal appointment time and the real optimal appointment time when the number of samples $n = 100, 1,000, 10,000$. The optimization objective is to minimize the $q$th quantiles of the total cost. The real optimal appointment time is acquired from the algorithm introduced in Section 2.1, while assuming the underlying distribution of the service duration is the log-normal distribution. We set the constant parameters as $\alpha = 2$, $\mu = 0$ and $\sigma = 1$.

**Remark 3.** *For robustness checks of our sampling approach, we run experiments with more instances: lognormal distribution with mean = 20 and standard deviation = 5 (Table 13), exponential distribution with mean = 20 (Table 14), and lognormal distribution with mean=20 and standard deviation=5 of with various combinations of cost coefficients (Table 15). We see that results hold and sampling approach results closely follow theoretical approach results. We provide these tables in the appendix.*

### *2.3. Unequal Costs Case*

In Section 2.1, we focus on the case where two cost coefficients $u_1$ and $o_1$ are equal. The basic techniques can be applied in the unequal cost coefficient case. The difficulty arises in obtaining the theoretical quantile function of the total cost in this scenario. For some distributions, such as the log-normal distribution, we may be able to find the theoretical quantile by using some algorithms, such as the Newton-Raphson method or algorithms proposed by [69], though there is no closed form of the theoretical quantile, even for the log-normal distribution. However, for some other distributions, even numerical computation of theoretical quantiles is not available and hence optimizing quantiles becomes a challenge. To address this difficulty, we employ the sampling approach developed in Section 2.2. As long as some independent observations of the service duration are available, the sample quantiles based on the empirical distribution of the service duration can be evaluated. As a consequence, the sample-based optimal appointment time can be obtained by minimizing the sample quantile of the total cost.

Next, we illustrate the basic idea of how to obtain the optimal appointment time numerically (without sampling) for a single job scheduling when the cost coefficients are unequal and the distribution of the service duration is known. Let the cumulative distribution function (cdf) and probability density function (pdf) of the service duration be denoted as $F_{U_1}$ and $f_{U_1}$, respectively.

**Proposition 3.** $F_{Y_1}$, the cdf of the total cost $Y_1$, is

$$F_{Y_1}(x) = \begin{cases} F_{U_1}(A_2 + \frac{x}{o_1}) & x \geq A_2 u_1 \\ F_{U_1}(A_2 + \frac{x}{o_1}) - F_{U_1}(A_2 - \frac{x}{u_1}) & \text{otherwise.} \end{cases}$$

*Proof.* The cumulative distribution function (cdf) of the total cost (denoted as $Y_1$) evaluated at $x$ ($x > 0$) is

$$\begin{aligned} a(x) &= P(Y_1 \leq x) \\ &= P(o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ \leq x) \\ &= P(o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ \leq x | C_1 \geq A_2) \times P(C_1 \geq A_2) \\ &+ P(o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ \leq x | C_1 < A_2) \times P(C_1 < A_2) \\ &= \frac{P(o_1(C_1 - A_2) \leq x, C_1 \geq A_2)}{P(C_1 \geq A_2)} P(C_1 \geq A_2) + \frac{P(u_1(A_2 - C_1) \leq x, C_1 < A_2)}{P(C_1 < A_2)} P(C_1 < A_2) \\ &= P(0 \leq o_1(C_1 - A_2) \leq x) + P(0 < u_1(A_2 - C_1) \leq x) \\ &= P\left(C_1 \leq A_2 + \frac{x}{o_1}\right) - P(C_1 \leq A_2) + P(C_1 \leq A_2) - P\left(C_1 \leq A_2 - \frac{x}{u_1}\right) \\ &= F_{U_1}\left(A_2 + \frac{x}{o_1}\right) - F_{U_1}\left(A_2 - \frac{x}{u_1}\right). \end{aligned} \tag{4}$$

In other words, $F_{Y_1}$, the cdf of the total cost $Y_1$, is

$$F_{Y_1}(x) = \begin{cases} F_{U_1}(A_2 + \frac{x}{o_1}) & x \geq A_2 u_1 \\ F_{U_1}(A_2 + \frac{x}{o_1}) - F_{U_1}(A_2 - \frac{x}{u_1}) & x < A_2 u_1 \end{cases}$$

□

**Remark**: The pdf $f_{Y_1}(x)$ ($x > 0$) can be expressed as

$$f_{Y_1}(x) = \begin{cases} \frac{f_{U_1}\left(A_2 + \frac{x}{o_1}\right)}{o_1} & \text{if } x \geq A_2 u_1 \\ \frac{f_{U_1}\left(A_2 + \frac{x}{o_1}\right)}{o_1} + \frac{f_{U_1}\left(A_2 - \frac{x}{u_1}\right)}{u_1} & \text{otherwise.} \end{cases}$$

Following the notation in Section (2.1), the $q$th quantile of $Y_1$, $Q_{Y_1}(A_2|q)$, satisfies

$$P(Y_1 \leq Q_{Y_1}(A_2|q)) = P(o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ \leq Q_{Y_1}(A_2|q)) = q.$$

With the above expression of cdf of $Y_1$, if $q \geq F_{U_1}\left(A_2 + \frac{A_2 u_1}{o_1}\right)$, then $Q_{Y_1}(A_2|q)$ is accessible if the $q$th quantile of the service duration is accessible, because $Q_{Y_1}(A_2|q) = (Q_{U_1}(q) - A_2) * o_1$ where $Q_{U_1}(q)$ denotes the $q$th quantile of the service duration. For those $q < F_{U_1}\left(A_2 + \frac{A_2 u_1}{o_1}\right)$, the Newton-Raphson type algorithms or the "uniroot" method based on [69] can be employed to find out the root of the following equation:

$$F_{U_1}\left(A_2 + \frac{x}{o_1}\right) - F_{U_1}\left(A_2 - \frac{x}{u_1}\right) = q.$$

The root of the above equation is the $q$th quantile of the total cost based on (4), which is denoted as $Q_{Y_1}(A_2|q)$. This root has a lower bound of 0 and upper bound of $A_2 u_1$, which can facilitate the searching of the root, which is unique.

Let $\hat{A}_2(q)$ denote the minimizer of $Q_{Y_1}(A_2|q)$. Now we derive a property which can be used to determine the range of the optimal appointment time $\hat{A}_2(q)$.

**Proposition 4.** Optimal appointment time $\hat{A}_2(q)$ and the corresponding $q$th quantile function of the total cost satisfy $Q_{Y_1}(\hat{A}_2(q)|q) < u_1 \hat{A}_2(q)$.

*Proof.* Since the support of $f$ is $\{x : x > 0\}$, i.e., $f(x) = 0$ if $x \leq 0$, the cdf of $Y_1$ can be written as

$$F_{Y_1}(x) = F_{U_1}\left(A_2 + \frac{x}{o_1}\right) - F_{U_1}\left(A_2 - \frac{x}{u_1}\right), \quad x > 0.$$

Thus, $Q_{Y_1}(A_2|q)$ satisfies

$$F_{U_1}\left(A_2 + \frac{Q_{Y_1}(A_2|q)}{o_1}\right) - F_{U_1}\left(A_2 - \frac{Q_{Y_1}(A_2|q)}{u_1}\right) = q. \tag{5}$$

Taking the derivative of (5) with respect to $A_2$ and evaluating it at $A_2 = \hat{A}_2(q)$, we get

$$f_{U_1}(\hat{A}_2(q) + \frac{Q_{Y_1}(\hat{A}_2(q)|q)}{o_1}) = f_{U_1}(\hat{A}_2(q) - \frac{Q_{Y_1}(\hat{A}_2(q)|q)}{u_1}). \tag{6}$$

From (6), we know that $Q_{Y_1}(\hat{A}_2(q)|q) < u_1 \hat{A}_2(q)$ since $\hat{A}_2(q) > 0$ and $f_{U_1}\left(\hat{A}_2(q) + \frac{Q_{Y_1}(\hat{A}_2(q)|q)}{o_1}\right) > 0$.

□

The result of Proposition 4 is useful to narrow down the search space where we look for the optimal appointment time.

With the unequal cost coefficients, we are unable to express $\hat{A}_2(q)$ as an average of two quantiles of the service duration as (3). When the theoretical median of the total cost is numerically inaccessible, our strategy is to look for $\hat{A}_2$, which minimizes the sample median of $Y_1$. This idea can be applied to any quantile of the total cost, and more importantly, no distribution assumption is needed.

Table 4 summarizes the optimal appointment time that minimizes the corresponding quantiles of the total cost with unequal coefficients of overage and underage costs. "T" stands for optimization based on the theoretical distribution of the total cost, while "S" stands for optimization based on the empirical distribution of the total cost. For the sampling method, the service duration is generated from a log-normal distribution with $\mu = 0$ and $\sigma = 1$, and the sample size is $n = 10,000$. With such a large sample size, there is little discrepancy between the theoretical method and the sampling method in terms of the optimal appointment time with different cost coefficients. In addition, the magnitude of the cost coefficients really matters. To be more specific, when the quantile to be optimized is fixed, changing the magnitude of the cost coefficients can affect the corresponding optimal appointment time.

| | $(o_1, u_1) = (1, 2)$ | | $(o_1, u_1) = (2, 1)$ | |
|---|---|---|---|---|
| | T | S | T | S |
| $q = 0.2$ | 0.35 | 0.34 | 0.45 | 0.44 |
| $q = 0.5$ | 0.44 | 0.45 | 0.75 | 0.76 |
| $q = 0.8$ | 0.82 | 0.83 | 1.58 | 1.58 |

Table 4: The optimal appointment time for $q$th quantiles of the total cost with unequal cost coefficients. The total cost $Y_1 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+$. "T" stands for optimization based on the theoretical distribution of the total cost; while "S" stands for optimization based on the empirical distribution of the total cost.

## 3. Multiple Jobs

In this section, we first study the case in which there are two jobs, and then discuss the case with three jobs or more.

### *3.1. Two Jobs*

We now consider the scenario in which two appointments need to be scheduled. As always, the first job starts at time zero; its completion time is given by $C_1 = U_1$ and the second job's completion time is given by $C_2 = \max\{A_2, C_1\} + U_2$. We represent the total cost for the two job appointment scheduling as

$$Y_2 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ + o_2(C_2 - A_3)^+ + u_2(A_3 - C_2)^+.$$

Our goal is to find the optimal appointment time $\hat{A}_2$ and $\hat{A}_3$, which minimizes the $q$th quantile of the total cost $Y_2$. We use the sample $q$th quantile instead of the population quantile as the objective function. This approach does not rely on the distribution of the service duration, and it can be applied as long as some independent observations of the service duration are available. The corresponding minimizer, $\hat{A}_2$ and $\hat{A}_3$, which minimizes the $q$th quantile of the sample is treated as the optimal appointment time. Let $\hat{Q}_{Y_2}(A_2, A_3|q)$ denote the sample $q$th quantile ($q * 100$th percentile) of the total cost $Y_2$. We choose to update $A_2$ and $A_3$ simultaneously to minimize $\hat{Q}_{Y_2}$. Next, we present numerical results using this optimization method. To implement the optimization, we use R's command **optim** which can accommodate non-differentiable functions. We do not assume any specific restrictions on cost coefficients $o_1, u_1, o_2, u_2$ and probability level $q$. For presentation clarity, we consider a case with $o_1 = u_1 = \alpha =$ and $o_2 = u_2 = \beta$. Then $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$. Without loss of generality, we assume that $q = 0.5$ (i.e., the median of the total cost is the objective function to be optimized) and use a log-normally distributed service duration distribution for the numerical results. We aim to find the values of $A_2$ and $A_3$, which minimize the median of $Y_2$.

Table 5 summarizes the estimated optimal appointment times $(\hat{A}_2, \hat{A}_3)$ based on the optimization approach introduced above with the setting $q = 0.5$, $\alpha = \beta = 1$, $\mu = 0$ and $\sigma = 1$. We use a sample size of $10^6$ for our dataset and use the regular median computation (by **median** in R). The optimization procedure is stable in terms of the initial value chosen in the algorithm.

| starting points $(A_2, A_3)$ | $\hat{A}_2$ | $\hat{A}_3$ | Time(s) |
|---|---|---|---|
| (1, 1.5) | 0.75 | 1.96 | 8.69 |
| (1, 2.5) | 0.75 | 1.96 | 8.67 |
| (3, 10) | 0.74 | 1.96 | 8.28 |

Table 5: Two-job optimization using the regular median computation with various starting values. $\hat{A}_2$ and $\hat{A}_3$ denote the optimal appointment time for the second and third job, respectively. "Time" refers to the time taken to conduct the optimization in seconds.

*3.2. Three or More Jobs*

The idea of two-job appointment scheduling can be extended to three or more jobs. Table 6 summarizes the optimal appointment times when there are three jobs, when the corresponding cost coefficients for each job are set as $o_1 = u_1 = 1,\ o_2 = u_2 = 2,\ o_3 = u_3 = 3$, $\mu = 0$, $\sigma = 1$ and the sample size is $N = 10^6$.

| starting points $(A_2, A_3, A_4)$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_4$ | Time(s) |
|---|---|---|---|---|
| (1, 1.5, 2) | 1.3 | 2.6 | 3.8 | 30.3 |
| (1, 2.5, 5) | 1.3 | 2.7 | 3.8 | 26.8 |
| (1, 5, 10) | 1.3 | 2.7 | 3.9 | 32.3 |

Table 6: Three-job optimization with various starting values. $\hat{A}_2$, $\hat{A}_3$, and $\hat{A}_4$ denote the optimal appointment time for the second, third, and fourth job, respectively. “Time” refers to the time taken to conduct the optimization in seconds.

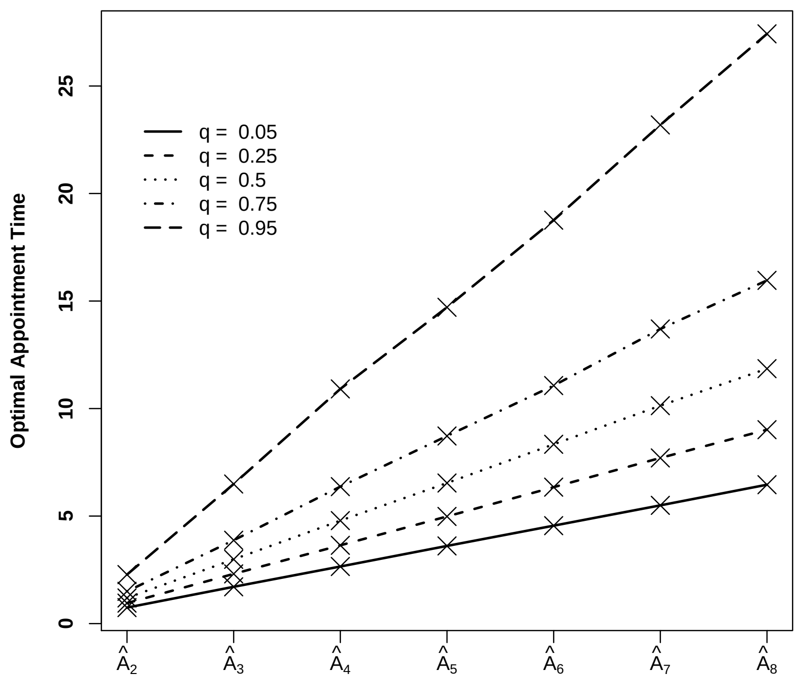



Figure 5: The optimal appointment time of seven jobs under different probability levels ($q$). Duration times are independently sampled from logNormal($\mu = 0, \sigma = 1$) and all cost coefficients are set as 1.

Next we consider a more complicated task, where there are seven jobs to schedule. The duration times of the seven jobs are independently and identically generated from logNormal ($\mu = 0$, $\sigma = 1$), and all cost coefficients are set to 1. The sample size is $n = 10^6$. Figure 5 depicts the optimal appointment times for different quantile ($q$) functions. We present a similar graph (Figure 13) in the appendix with lognormal distribution with a mean of 20 and standard deviation of 5. We find that the optimal appointment time for each job displays almost like linear shape, regardless of which quantile function of the total cost is the target. Furthermore, the gap between the optimal times to two consecutive jobs increases as the quantile level ($q$) increases, i.e., if the target function of interest is the high level quantile function of the total cost, then the optimal appointment time for each job is considerably more widespread compared with low level quantiles, when all cost coefficients are equal and the duration times are identically distributed.

It is a well-known result that in the case of expected cost criteria and identical jobs (same durations and same cost coefficients), optimal allocated appointment times (optimal allocated durations) exhibit a dome-shaped pattern, i.e., the optimal durations between two consecutive jobs are increasing initially but then decreasing later ([23, 8]). Our results show that when the objective is minimizing a quantile of the total cost, the optimal appointment times do not show the well-known dome-shaped pattern but they exhibit a linear shape regardless of the quantile level as in Figure 5. When we consider the allocated appointment times (durations), they exhibit a semi-dome-shaped pattern which first increases (like the well-known dome-shaped pattern) but then its decrease is not monotone (unlike the well-know dome-shaped pattern) as shown in Figure 6.

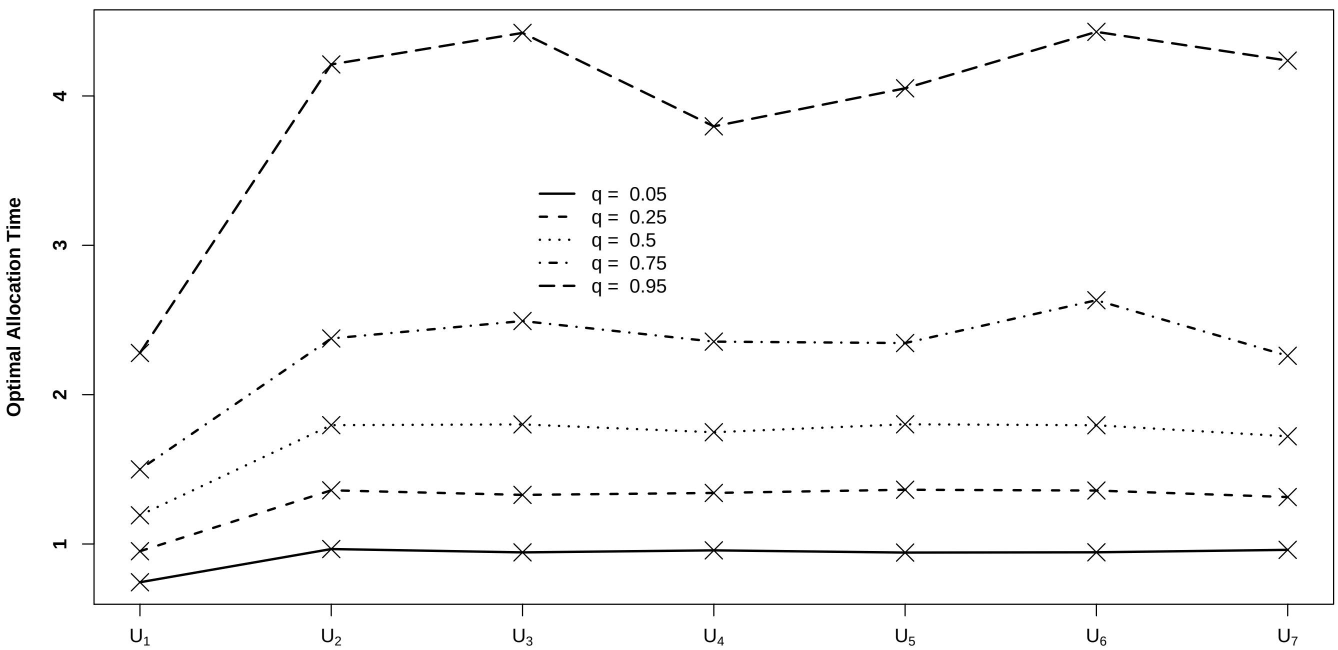



Figure 6: The optimal allocation time of seven jobs under different quantile levels ($q$). Duration times are independently sampled from logNormal ($\mu = 0, \sigma = 1$) and all cost coefficients are set as 1.

We also investigate how the schedule changes when one uses different job sequence as suggested by one of the referees. For this experiment, we use some of the well-known sequencing rules from the literature. We have four jobs to be scheduled, and the duration times follow lognormal distributions with ($\mu = 0, \sigma = 1$), ($\mu = 1, \sigma = 1$), ($\mu = 0, \sigma = 2$) and ($\mu = 1, \sigma = 2$), respectively. Additionally, for each of them, we take that the overage cost is equal to the underage cost and the overage costs are 1, 2, 4 and 16, respectively. We consider three sequencing rules: (1) sequence jobs with increasing variance (rule 1) [3], i.e., schedule smallest variance job first and highest variance job the last; (2) sequence jobs with increasing order of coefficient of variation (CV) [3], i.e., schedule the smallest of CV first and schedule the highest CV the last, and this would be the same with rule 1 in our experiment setup; and (3) schedule jobs with increasing order of the ratio of standard deviation/overage cost (rule 3) [39], i.e., schedule the smallest ratio of standard deviation/overage cost first. The optimal appointment times corresponding to these three rules are displayed in Figure 14. As expected, we see that different sequencing rules give us different appointment times. We find that at the same quantile level ($q$), the third sequencing rule leads to the lowest resulting cost, which is the $q$th quantile function of the cost evaluated at the optimal appointment time. Also please note for this particular example the first and second sequencing rules yield the same sequencing of the four jobs, and that is why their appointment times are same.

### 3.3. Computation efficiency

The computational efficiency of our proposed algorithm depends on the number of jobs and the sample size. We now investigate the computation time of our algorithm and its relation to the number of jobs and the sample size.

| | $n = 100$ | $n = 10000$ | $n = 1000000$ |
|---|---|---|---|
| $N = 3$ | 0.035 | 0.142 | 17.115 |
| $N = 5$ | 0.033 | 0.332 | 49.406 |
| $N = 7$ | 0.149 | 0.598 | 80.813 |
| $N = 9$ | 0.218 | 0.890 | 202.221 |

Table 7: Computation time (in seconds) when minimizing the median ($q = 0.5$) of the total cost with different combinations of the number of jobs (denoted by $N$) and sample size (denoted by $n$).

Following the set up of Section 3.2, we generated $N$ jobs with duration times of i.i.d logNormal ($\mu = 0$, $\sigma = 1$) and all cost coefficients are set to 1. Table 7 summarizes the computational time of our algorithm to minimize the median ($q = 0.5$) of the total cost for $N$ jobs with different sample sizes $n$. We ran these computation experiments on a laptop with an Apple M1 chip (3.2 GHz and 16-core Neural Engine) and 16GB RAM. Obviously, more computational time is needed to complete the optimization procedure as sample size increases. This is not surprising since most of the computational time is spent on evaluating the sample quantile of the total cost, especially when the sample size is large, e.g., $10^6$. If we reduce the sample size to the order of $10^4$, it only takes 0.89 seconds to schedule 9 jobs in comparison with 203 seconds (still less than four minutes) with $n = 10^6$. These findings suggest that the sample size plays a more important role than the number of jobs to be scheduled in determining computational efficiency. Finding a more efficient way to evaluate sample quantile may be the solution to improve the efficiency of our algorithm with large sample sizes. For example, [70] introduced a fast approach to calculate the sample median and use of this method can speed up our algorithm. We are planning to implement this fast method to compute sample median as well as other approaches to improve the efficiency of the algorithm in future work.

## 4. Numerical Examples with Real Surgery Data

We demonstrate our method with numerical examples based on real data for surgery scheduling. The data consists of surgical durations, and it was compiled from multiple Canadian hospitals. Each data point corresponds to a surgery and consists of two fields: 1) procedure length (in minutes) and 2) type (in 2 or 3 character codes). The type of surgery is determined and assigned by healthcare providers, and it may be at the procedure level or at a higher level. The data contains about 70 different types and over 40,000 surgeries.

We present optimal appointment allocations and the corresponding quantile costs with non-identical jobs, identical jobs and a single job. We compare quantile solutions with the ones using expected cost criterion. We show that the optimal scheduling time has a sizeable differences between both of these two criteria. We start with non-identical jobs.

### *4.1. Non-Identical Jobs*

In this subsection, we provide numerical examples and results for two non-identical job scheduling cases. We minimize the quantile of the total cost for a variety of cost coefficients and probability levels. We choose the Type C3 and C4 jobs as the first and second job, respectively. The data set has 358 jobs in Type C3 and 240 jobs in Type C4. The mean and the standard deviation of Type C3 jobs are 143.33 minutes and 135.85 minutes, respectively; while those of Type C4 are 133.11 minutes and 179.61 minutes, respectively. We randomly select 10,000 pairs of service durations from all the possible combinations of C3 and C4 jobs. These 10,000 pairs of service durations are used for the analysis.

Table 8 summarizes the optimal appointment time under different combinations of cost coefficients based on optimizing the quantile function of the total cost. Here, we assume that for each job, the underage cost coefficient is equal to the overage cost coefficient, i.e, the total cost is $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$. $\hat{A}_2$ and $\hat{A}_3$ denote the optimal appointment time for the second and third job, respectively. Moreover, Table 16 in the appendix summarizes the minimal quantile functions of the total cost under different combinations of the cost coefficients and the probability level ($q$), when the optimal appointment time shown in Table 8 is scheduled.

| $(\alpha, \beta)$ | $q = 0.05$ | | $q = 0.25$ | | $q = 0.50$ | | $q = 0.75$ | | $q = 0.95$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ |
| (9, 1) | 56.0 | 106.0 | 51.4 | 130.4 | 72.0 | 197.1 | 117.1 | 275.5 | 194.7 | 458.7 |
| (8, 2) | 55.8 | 98.8 | 46.7 | 118.5 | 80.4 | 187.9 | 109.1 | 284.0 | 175.7 | 521.1 |
| (7, 3) | 52.9 | 99.2 | 47.2 | 117.2 | 82.6 | 192.4 | 124.5 | 276.3 | 181.5 | 501.8 |
| (6, 4) | 59.5 | 104.5 | 46.9 | 112.6 | 86.6 | 190.1 | 126.9 | 267.1 | 149.3 | 498.8 |
| (5, 5) | 58.0 | 107.5 | 45.3 | 110.3 | 101.7 | 187.2 | 154.0 | 263.5 | 153.7 | 479.7 |
| (4, 6) | 52.0 | 99.0 | 51.3 | 113.1 | 120.3 | 193.8 | 169.3 | 275.6 | 175.0 | 474.2 |
| (3, 7) | 50.8 | 98.1 | 58.3 | 119.0 | 138.0 | 207.5 | 194.1 | 289.2 | 184.5 | 481.4 |
| (2, 8) | 55.0 | 106.5 | 86.0 | 133.6 | 167.2 | 229.7 | 249.3 | 332.1 | 189.8 | 487.5 |
| (1, 9) | 52.0 | 103.9 | 178.5 | 224.1 | 212.9 | 274.6 | 301.9 | 380.4 | 211.3 | 509.4 |

Table 8: Summary of the two-job optimal appointment times with various combinations of cost coefficients and probability levels ($q$). The total cost $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$.

We also compare the optimal appointment time minimizing the sample quantile function of the total cost with that minimizing the expected total cost. Table 9 gives the optimal appointment times when the median of the total cost and the expected total cost are optimized with different combinations of cost coefficients. $\hat{A}_2$ and $\hat{A}_3$ denote the optimal appointment time for the second and third job, respectively. $Q_{Y_2}$ denotes the corresponding median or mean of the total cost under the optimal appointment time.

| $(\alpha, \beta)$ | Median | | | Mean | | |
|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $Q_{Y_2}$ | $\hat{A}_2$ | $\hat{A}_3$ | mean |
| (9, 1) | 72.0 | 197.1 | 633.5 | 124.0 | 246.0 | 1007.2 |
| (8, 2) | 80.4 | 187.9 | 654.6 | 134.3 | 251.3 | 1038.8 |
| (7, 3) | 82.6 | 192.4 | 654.6 | 137.0 | 253.0 | 1068.4 |
| (6, 4) | 86.6 | 190.1 | 649.8 | 147.0 | 258.0 | 1095.7 |
| (5, 5) | 101.7 | 187.2 | 645.8 | 159.0 | 264.0 | 1119.2 |
| (4, 6) | 120.3 | 193.8 | 626.1 | 172.0 | 271.0 | 1137.8 |
| (3, 7) | 138.0 | 207.5 | 596.5 | 191.0 | 284.0 | 1148.4 |
| (2, 8) | 167.2 | 229.7 | 546.3 | 219.0 | 307.0 | 1146.8 |
| (1, 9) | 212.9 | 274.6 | 464.4 | 282.0 | 357.0 | 1116.9 |

Table 9: Two-job optimal appointment times with respect to median and expected cost objectives with different combinations of cost coefficients. The total cost $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$. $Q_{Y_2}$ denotes the corresponding median of the total cost under the optimal appointment time.

The results show that allocated appointment times under median cost and expected cost criteria are quite different. The optimal appointment time minimizing expected total cost is larger than the optimal appointment time minimizing lower to middle quantiles of the total cost. When minimizing the expected total cost, some extremely large values of the total cost need to be accounted for, which forces the corresponding optimal appointment time to be larger. On the other hand, if lower to middle quantiles of the total cost are to be optimized, we just need to account for those small to moderate total values of the total cost. This fact might lead to the great gap in the optimal appointment time when the median and expected total cost are to be minimized.

Following a referee's suggestion, we also consider appointment allocation with unequal underage and overage cost efficients. We take that uniform underage/overage cost coefficients across jobs, i.e., $o_1 = o_2$ and $u_1 = u_2$ but $o_1 \neq u_1$, with the jobs to be scheduled as defined above. The corresponding optimal appointment times, the minimal quantile functions of the total cost and optimal appointment times to minimize the median and/or the mean of the total cost were displayed in Tables 17, 18 and 19, respectively, in the appendix.

In addition to two-job appointment, we further carry out scheduling three jobs with our proposed method for another referee suggestion. The data set has 859 jobs in Type C5, whose mean and standard deviation are 95.57 and 56.12 minutes, respectively. We randomly select 10,000 pairs of service durations from all the possible combinations of C3, C4 and C5 jobs. These 10,000 pairs of service durations are used for the analysis. Here we adopt the first two sequencing rules aforementioned to determine the order of C3, C4 and C5; that is, C5 $\rightarrow$ C3 $\rightarrow$ C4. Again we take different underage and overage cost coefficients for each job: The total cost is defined by $Y_3 = \sum_{i=1}^{3} \{o_1(C_i - A_{i+1})^+ + u_1(A_{i+1} - C_i)^+\}$. Table 10 displays the corresponding optimal appointment times under different combinations of the cost coefficients and the probability level ($q$).

| $(o_1, u_1)$ | $q = 0.05$ | | | $q = 0.25$ | | | $q = 0.75$ | | | $q = 0.95$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_4$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_4$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_4$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_4$ |
| (9, 1) | 55.8 | 115.1 | 182.3 | 108.7 | 204.7 | 315.7 | 191.6 | 552.9 | 846.7 | 548.4 | 1030.3 | 1922.0 |
| (8, 2) | 554.1 | 113.4 | 171.2 | 95.8 | 183.3 | 277.7 | 160.0 | 491.6 | 696.1 | 491.8 | 873.0 | 1482.6 |
| (7, 3) | 50.3 | 109.6 | 166.9 | 84.9 | 164.7 | 249.2 | 124.4 | 348.3 | 575.6 | 352.8 | 751.6 | 1263.7 |
| (6, 4) | 46.7 | 103.7 | 156.5 | 68.5 | 141.1 | 226.6 | 102.0 | 315.5 | 505.1 | 301.8 | 634.9 | 1046.3 |
| (5, 5) | 43.3 | 99.7 | 147.7 | 54.9 | 133.5 | 210.5 | 71.2 | 287.3 | 433.5 | 206.5 | 475.2 | 862.6 |
| (4, 6) | 34.0 | 86.3 | 136.0 | 46.6 | 116.8 | 187.0 | 56.3 | 221.1 | 372.0 | 177.2 | 401.7 | 711.3 |
| (3, 7) | 35.4 | 88.1 | 135.9 | 39.3 | 103.9 | 175.6 | 47.9 | 196.3 | 308.3 | 130.8 | 332.9 | 566.1 |
| (2, 8) | 40.3 | 102.6 | 159.4 | 35.7 | 97.7 | 155.0 | 43.9 | 164.0 | 235.8 | 106.4 | 237.4 | 411.4 |
| (1, 9) | 45.4 | 108.9 | 151.1 | 29.9 | 75.5 | 133.9 | 30.0 | 105.2 | 166.2 | 57.0 | 137.6 | 241.9 |

Table 10: Summary of the three-job optimal appointment times with various combinations of cost coefficients and probability levels.

We also compared the optimal appointment allocations under median cost and expected cost criteria in this three-job setting.

| $(o_1, u_1)$ | Median | | | | Mean | | | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_4$ | $Q_{Y_3}$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_4$ | $Q_{Y_3}$ |
| (9, 1) | 123.6 | 356.5 | 481.6 | 377.6 | 462.1 | 829.1 | 1212.0 | 1309.0 |
| (8, 2) | 104.8 | 301.1 | 413.5 | 637.0 | 277.0 | 571.0 | 797.0 | 1823.6 |
| (7, 3) | 88.4 | 250.7 | 367.3 | 813.8 | 176.0 | 440.0 | 606.0 | 2036.6 |
| (6, 4) | 73.5 | 230.5 | 326.3 | 921.4 | 125.0 | 356.0 | 485.0 | 2059.9 |
| (5, 5) | 57.3 | 200.9 | 291.9 | 979.6 | 93.0 | 295.0 | 404.1 | 1959.3 |
| (4, 6) | 49.7 | 174.8 | 257.2 | 973.3 | 70.0 | 250.0 | 343.0 | 1766.6 |
| (3, 7) | 41.8 | 149.8 | 220.0 | 889.2 | 54.0 | 200.0 | 286.0 | 1482.3 |
| (2, 8) | 34.4 | 120.3 | 192.0 | 712.4 | 42.0 | 147.0 | 219.0 | 1100.5 |
| (1, 9) | 25.9 | 93.7 | 153.9 | 430.4 | 31.0 | 97.4 | 158.0 | 612.2 |

Table 11: Three-job optimal appointment times with respect to median and expected cost objectives with different combinations of cost coefficients. $Q_{Y_3}$ denotes the corresponding median or mean of the total cost under the optimal appointment time.

### *4.2. Identical Jobs*

Now consider two identical jobs; that is, jobs that have identical duration distributions. We chose type H2 surgery from our data set, and there are 10,835 observations. We randomly chose 5,409 pairs from all the possible pairs of observations and used them in the analysis.

Table 20 in the appendix summarizes the optimal appointment time under different combinations of cost coefficients when minimizing varying quantiles of the total cost. The combinations of cost coefficients and probability levels chosen here are the same as those used in Section 4.1. Additionally, Table 21 in the appendix compares the quantiles of the total cost for the corresponding optimal times in Table 20. An interesting observation for the identical two-job case is that the optimal appointment time is robust to the choice of cost coefficients, especially for the low to moderate quantile functions, $q < 0.75$. When minimizing 0.95th quantile function of the total cost, there is an abrupt increase in the optimal appointment time, in contrast to low to moderate quantile functions.

Table 22 in the appendix gives identical two-job optimal appointment times for median and expected cost optimization with different combinations of cost coefficients. Similar to what we observe for the non-identical case, the appointment time that minimizes the expected total cost is larger than the optimal appointment time that minimizes the lower to middle quantiles of the total cost. Furthermore, the corresponding minimal expected total cost is larger than the median, evaluated at the optimal appointment time, of the total cost.

In the next subsection, we present results for a single job in which we compare the quantile solution with the mean cost solution for a variety of cost coefficients and quantile levels.

### *4.3. Single Job*

For single job scheduling, we chose service duration of type C4 as a candidate, since a quantile-quantile plot and Kolmogorov-Smirnov test showed that log-normal distribution, log-normal($\mu$, $\sigma^2$) is a good fit for the data. This approach allows us to make a comparison between minimizing the theoretical and empirical quantiles of the total cost. The sample size is 237.

The two unknown parameters, $\mu$ and $\sigma$, in the log-normal distribution are replaced by their maximum likelihood estimators (MLEs), respectively. As shown in Section 2.3, whether the two cost coefficients are equal or not, the cdf and pdf of the total cost are analytically tractable when the distribution of service duration is known. Moreover, the quantile function is numerically tractable, see Remark 4. Therefore, we are able to minimize the corresponding quantile function of the total cost with respect to the appointment time, which is denoted as $A_2$. On the other hand, employing the sampling method used in Section 2.2, we can minimize the quantile based on the empirical distribution of the total cost as well.

**Remark 4.** *The single job quantile function is numerically tractable in the case of lognormal distribution. To see this we can use the result of Proposition 2 to find the quantile function. In Proposition 2, we established the relationship between the cumulative distribution of the total cost $Y_1$ and that of $U_1$, which follows a log-normal distribution. Now let's the derive the qth quantile of $Y_1$.*

*If $q > F_{U_1}(A_2 + A_2U_1/o_1)$, then $F_{Y_1}(x) = q$ if and only if $F_{U_1}(A_2 + x/o_1) = q$. In other words,*

$$\begin{aligned} q &= P(U_1 \le A_2 + x/o_1) \\ &= P(\log(U_1) \le \log(A_2 + x/o_1) \\ &= P\left(\frac{\log(U_1) - \mu}{\sigma} \le \frac{\log(A_2 + x/o_1) - \mu}{\sigma}\right) \end{aligned}$$

*Therefore, $\frac{\log(A_2+x/o_1)-\mu}{\sigma} = \Phi^{-1}(q)$, which implies that $x = o_1(\exp(\mu + \sigma\Phi^{-1}(q)) - A_2)$. On the other hand, if $q \le F_{U_1}(A_2 + A_2U_1/o_1)$, we need to find the root of the equation*

$$F_{U_1}(A_2 + x/o_1) - F_{U_1}(A_2 - x/o_1) = q$$

*where $U_1$ follows log-normal $(\mu, \sigma)$. Numerical methods such as a Newton type algorithm or the bisection method can be employed to find the root.*

Table 12 summarizes optimal appointment time with different combinations of cost coefficients based on minimizing quantiles of the total cost. In the table, for comparison, the last row shows the corresponding optimal appointment time when expected total cost is to be minimized under these cost coefficients combinations. Here, five candidates of quantiles and nine combinations of cost coefficients are considered. For each combination of cost coefficients and probability level ($q$), the upper row, denoted by "P", shows the optimal appointment time assuming the service duration follows a log-normal distribution, while the lower row, denoted by "NP", shows the optimal appointment time based on the sampling method, as described in Section 2.3.

| $(o_1, u_1)$ | | (9, 1) | (8, 2) | (7, 3) | (6, 4) | (5, 5) | (4, 6) | (3, 7) | (2, 8) | (1, 9) |
|---|---|---|---|---|---|---|---|---|---|---|
| $q = 0.05$ | P | 35.8 | 35.1 | 34.6 | 34.0 | 33.4 | 32.8 | 32.3 | 31.7 | 31.1 |
| | NP | 71.5 | 42.4 | 69.5 | 40.4 | 50.0 | 48.6 | 89.8 | 91.2 | 88.6 |
| $q = 0.25$ | P | 48.5 | 45.5 | 42.5 | 39.5 | 36.5 | 33.5 | 30.5 | 27.5 | 24.5 |
| | NP | 50.8 | 40.0 | 45.1 | 45.2 | 43.0 | 40.8 | 37.6 | 36.4 | 31.3 |
| $q = 0.50$ | P | 74.9 | 68.1 | 61.2 | 54.4 | 47.6 | 40.8 | 34.0 | 27.2 | 20.4 |
| | NP | 66.7 | 61.4 | 56.1 | 50.8 | 45.5 | 40.2 | 34.9 | 29.6 | 24.3 |
| $q = 0.75$ | P | 130.2 | 116.6 | 103.0 | 89.4 | 75.8 | 62.2 | 48.5 | 34.9 | 21.3 |
| | NP | 117.5 | 106.0 | 94.5 | 83.0 | 71.5 | 60.0 | 46.8 | 37.0 | 25.5 |
| $q = 0.95$ | P | 309.3 | 275.3 | 241.3 | 207.3 | 173.3 | 139.2 | 105.2 | 71.2 | 37.2 |
| | NP | 421.5 | 370.2 | 330.5 | 285.0 | 234.0 | 194.0 | 143.2 | 103.0 | 57.5 |
| Mean | P | 245.9 | 164.7 | 123.3 | 96.3 | 76.4 | 60.7 | 47.4 | 35.5 | 23.7 |
| | NP | 280.0 | 153.0 | 111.0 | 90.0 | 68.0 | 55.0 | 47.0 | 35.0 | 26.0 |

Table 12: Optimal appointment times that minimize the quantiles and the expected of the total cost with various combinations of cost coefficients. The total cost $Y_1 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+$. Here "P" and "NP" stand for the optimal appointment time with a parametric (log-normal) and a non-parametric (empirical) distribution of the duration time, respectively.

We see that when assuming that the service duration is log-normally distributed and the objective function ($q$th quantile of the total cost) is fixed - increasing underage cost coefficients while decreasing overage cost coefficients and keeping their sum fixed - the optimal appointment time will decrease. A possible interpretation is that increasing underage cost coefficients would result in a larger proportion of the cost induced by doctors' idling being reflected in the total cost. Thus, decreasing the appointment time may mitigate this idling cost to some extent. The same property applies to minimizing the mean of the total cost. On the other hand, when the combination of cost coefficients is fixed and the overage cost coefficient $o_1$ is greater than the underage cost coefficient $u_1$, the corresponding optimal appointment time increases as the quantile level $q$ increases. A possible interpretation is that when $o_1 > u_1$, particularly when there is a big gap between them, the cost induced by patients' waiting might dominate the other cost induced by idling. Hence, to minimize the larger quantiles of the total cost, we need to put off the appointment to reduce the cost induced by patients' waiting. In addition, compared with the optimal appointment time minimizing the mean of the total cost, the optimal appointment time that minimizes quantiles (except extremely large quantiles) of the total cost, seems to be more stable under different combinations of cost coefficients. In other words, minimizing the expected cost is more sensitive to the cost coefficients compared with minimizing quantiles of the total cost. This property poses a challenge for minimizing the expected cost, since determining cost coefficients appropriately is usually difficult in most real world situations.

As a comparison, some properties summarized above may apply to the case where there is no distributional assumption with respect to the service duration and the optimal appointment time is obtained from the sampling method. However, we could also find some exceptions when $q$ is very large or small. To be more specific, take $q = 0.05$ as an example. We observe a strictly monotone decreasing optimal appointment times as $o_1$ decreases and $u_1$ increases when service durations follow log-normal distribution. On the other hand, the trend for optimal appointment times is much more complicated when there is no distributional assumption for the service duration. A possible reason might be that the extremely low sample quantiles are prone to being affected by a few observations of the service duration compared with theoretical quantiles of the total cost when assuming service duration follows log-normal distribution. This may also explain why the corresponding sample-based optimal appointment time, which minimizes the $q$th quantile of the total cost when cost coefficients are fixed, is not strictly monotone increasing as $q$ increases, although an increasing pattern is observed. When extreme quantile functions (such as 5% or 95%) of the total cost are to be minimized, a greater discrepancy is presented between the optimal appointment time when assuming the log-normal distribution for the service duration, and that when there is no distributional assumption for the service duration. For other moderate quantile functions, the discrepancies are relatively small. A possible reason might be that extreme quantile functions of an empirical distribution are prone to being affected by extremely large or small values, compared with those based on a parametric distribution.

## 5. Conclusion

We determine an optimal appointment schedule that minimizes a quantile of the total (idle time, overtime and wait) cost for a given job sequence on a single server when server durations are random. Appointment scheduling has many important applications from a variety of industries with high economic and operational impact. To the best of our knowledge, our paper is the first that has a measure of quantile (of the total cost) for the objective function for appointment scheduling problem. This measure is directly related to service levels, and for some of the applications it may be more appropriate than the expected total cost.

Our approach is flexible and it does not necessarily minimize the worst case scenario like a robust approach would. Compared to robust approaches, our method is less conservative and it allows the decision maker to choose the level of the quantile cost to be minimized. Furthermore, we do not impose any conditions on the cost coefficients and therefore the decision maker can choose to minimize the wait time only, overtime only, or any desired combination of idle time, wait time and overtime.

We first analyze a special case of the general problem (a single job with equal cost coefficients and log normal processing distribution), obtain insights and extend our results to a greater number of jobs, generic coefficients and any distribution. We also allow any percentile level; that is, we do not focus on the median only. We obtain theoretical results for some special cases and develop methods for the general case. Our methods do not require a specific distribution and can work with samples. We present numerical examples with real data on surgeries. Results show that the schedules obtained under the new objective can be quite different than the ones when the expected cost objective is used.

Future work can study the effect of no-shows and walk-ins in the schedule under this new objective. Another interesting extension would be to consider this new quantile objective for advance scheduling problems and combined scheduling problems (which considers both advance and appointment scheduling problems, e.g., see [61, 60]). We expect that new solution approaches (such as decomposition methods, e.g., see ([71, 72, 46]) are needed to address these suggested future work since they are more complex and involve multiple levels of decision making.

## Acknowledgements

This research was supported by a discovery grant of J. Cao from the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank the reviewing team for their feedback which has significantly improved our paper.

## References

[1] D. Gupta, B. Denton, Health care appointment systems: Challenges and opportunities, IIE Transactions 40 (2008) 800–819.

[2] E. Demeulemeester, J. Beliën, B. Cardoen, M. Samudra, Operating room planning and scheduling, in: Handbook of healthcare operations management, Springer, 2013, pp. 121–152.

[3] B. Denton, J. Viapiano, A. Vogl, Optimization of surgery sequencing and scheduling decisions under uncertainty, Health care management science 10 (1) (2007) 13–24.

[4] A. Macario, Are your hospital operating rooms efficient? A scoring system with eight performance indicators, The Journal of the American Society of Anesthesiologists 105 (2) (2006) 237–240.

[5] J. A. Girotto, P. F. Koltz, G. Drugas, Optimizing your operating room: Or, why large, traditional hospitals don't work, International Journal of Surgery 8 (5) (2010) 359–367.

[6] S. Batun, M. A. Begen, Optimization in healthcare delivery modeling: Methods and applications, in: Handbook of Healthcare Operations Management, Springer, 2013, pp. 75–119.

[7] Canadian Institute for Health Information, Health spending in Canada 2019 (2019).
URL `https://www.cihi.ca/en/health-spending`

[8] L. W. Robinson, R. R. Chen, Scheduling doctors' appointments: Optimal and empirically-based heuristic policies, IIE Transactions 35 (2003) 295–307.

[9] J. Patrick, M. L. Puterman, M. Queyranne, Dynamic multipriority patient scheduling for a diagnostic resource, Operations Research 56 (2008) 1507 – 1525.

[10] A. Saure, J. Patrick, S. Tyldesley, M. L. Puterman, Dynamic multi-appointment patient scheduling for radiation therapy, European Journal of Operational Research 223 (2) (2012) 573–584.

[11] F. Sabria, C. F. Daganzo, Approximate expressions for queuing systems with scheduling arrivals and established service order, Transportation Science 23 (1989) 159–165.

[12] I. Bendavid, B. Golany, Setting gates for activities in the stochastic project scheduling problem through the cross entropy methodology, Annals of Operations Research 189 (2011) 25–42.

[13] M. Elhafsi, Optimal leadtime planning in serial production systems with earliness and tardiness costs, IIE Transactions 34 (2002) 233 – 243.

[14] I. Gurvich, J. Luedtke, T. Tezcan, Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach, Management Science 56 (7) (2010) 1093–1115.

[15] S. Shen, J. Wang, Stochastic modeling and approaches for managing energy footprints in cloud computing service, Service Science 6 (1) (2014) 15–33.

[16] P. P. Wang, Static and dynamic scheduling of customer arrivals to a single-server system, Naval Research Logistics 40 (3) (1993) 345–360.

[17] P. M. V. Bosch, D. C. Dietz, J. R. Simeoni, Scheduling customer arrivals to a stochastic service system, Naval Research Logistics 46 (1999) 549–559.

[18] T. Cayirli, E. Veral, Outpatient scheduling in health care: A review of literature, Production and Operations Management 12 (4) (2003) 519–549.

[19] C. Zacharias, T. Yunes, Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs, Management Science (2019).

[20] B. P. Berg, B. T. Denton, S. A. Erdogan, T. Rohleder, T. Huschka, Optimal booking and scheduling in outpatient procedure centers, Computers & Operations Research 50 (2014) 24–37.

[21] T. Cayirli, E. Veral, H. Rosen, Designing appointment scheduling systems for ambulatory care services, Health care management science 9 (1) (2006) 47–58.

[22] M. A. Begen, M. Queyranne, Appointment scheduling with discrete random durations, Mathematics of Operations Research 36 (2) (2011) 240–257.

[23] B. Denton, D. Gupta, A sequential bounding approach for optimal appointment scheduling, IIE Transactions 35 (2003) 1003–1016.

[24] G. C. Kaandorp, G. Koole, Optimal outpatient appointment scheduling, Health Care Management Science 10 (2007) 217–229.

[25] P. M. V. Bosch, D. C. Dietz, Minimizing expected waiting in a medical appointment system, Iie Transactions 32 (9) (2000) 841–848.

[26] T. Cayirli, K. K. Yang, S. A. Quek, A universal appointment rule in the presence of no-shows and walk-ins, Production and Operations Management 21 (4) (2012) 682–697.

[27] L. W. Robinson, Y. Gerchak, D. Gupta, Appointment times which minimize waiting and facility idleness, Working Paper, DeGroote School of Business, McMaster University (1996).

[28] E. N. Weiss, Models for determining estimated start times and case orderings in hospital operating rooms, IIE Transactions 22 (2) (1990) 143–150.

[29] T. R. Rohleder, K. J. Klassen, Rolling horizon appointment scheduling: A simulation study, Health Care Management Science 5 (3) (2002) 201–209.

[30] C.-J. Ho, H.-S. Lau, Minimizing total cost in scheduling outpatient appointments, Management science 38 (12) (1992) 1750–1764.

[31] K. J. Klassen, T. R. Rohleder, Scheduling outpatient appointments in a dynamic environment, Journal of Operations Management 14 (2) (1996) 83–101.

[32] L. R. LaGanga, S. R. Lawrence, Clinic overbooking to improve patient access and increase provider productivity, Decision Sciences 38 (2) (2007) 251–276.

[33] D. Shmoys, S. Henderson, C. Tong, Real-time pooling for multi-site imaging facilities, Working Paper, Cornell University (2015).

[34] K. J. Klassen, R. Yoogalingam, Improving performance in outpatient appointment services with a simulation optimization approach, Production and Operations Management 18 (4) (2009) 447–458.

[35] Z. Zhang, X. Xie, Simulation-based optimization for surgery appointment scheduling of multiple operating rooms, IIE Transactions 47 (2015) 998–1012.

[36] H. J. Schuetz, R. Kolisch, Approximate dynamic programming for capacity allocation in the service industry, European Journal of Operational Research 218 (1) (2012) 239–250.

[37] L. Green, S. Savin, B. Wang, Managing patient service in a diagnostic medical facility, Operations Research 54 (2006) 11–25.

[38] S. Choi, W. E. Wilhelm, Sequencing in an appointment system with deterministic arrivals and non-identical exponential service times, Computers & Operations Research 117 (2020) 104901.

[39] H. Y. Mak, Y. Rong, J. Zhang, Appointment scheduling with limited distributional information, Management Science 61 (2) (2014) 316–334.

[40] M. A. Begen, R. Levi, M. Queyranne, Technical note - a sampling-based approach to appointment scheduling, Operations research 60 (3) (2012) 675–681.

[41] Q. Kong, C.-Y. Lee, C.-P. Teo, Z. Zheng, Scheduling arrivals to a stochastic service delivery system using copositive cones, Operations research 61 (3) (2013) 711–726.

[42] D. Ge, G. Wan, Z. Wang, J. Zhang, A note on appointment scheduling with piecewise linear cost functions, Mathematics of Operations Research 39 (4) (2013) 1244–1251.

[43] S. Rachuba, B. Werners, A robust approach for scheduling in hospitals using multiple objectives, Journal of the Operational Research Society 65 (4) (2013) 546–556.

[44] P. Santibáñez, M. Begen, D. Atkins, Surgical block scheduling in a system of hospitals: an application to resource and wait list management in a British Columbia health authority, Health care management science 10 (3) (2007) 269–282.

[45] B. T. Denton, A. J. Miller, H. J. Balasubramanian, T. R. Huschka, Optimal allocation of surgery blocks to operating rooms under uncertainty, Operations research 58 (4-part-1) (2010) 802–816.

[46] B. Naderi, V. Roshanaei, M. A. Begen, D. M. Aleman, D. R. Urbach, Increased surgical capacity without additional resources: Generalized operating room planning and scheduling, to appear in Production & Operations Management (2021).

[47] M. Hamid, M. M. Nasiri, F. Werner, F. Sheikhahmadi, M. Zhalechian, Operating room scheduling by considering the decision-making styles of surgical team members: a comprehensive approach, Computers & Operations Research 108 (2019) 166–181.

[48] A. Najjarbashi, G. J. Lim, A variability reduction method for the operating room scheduling problem under uncertainty using cvar, Operations Research for Health Care 20 (2019) 25–32.

[49] D. Bertsimas, D. B. Brown, C. Caramanis, Theory and applications of robust optimization, SIAM review 53 (3) (2011) 464–501.

[50] A. Ben-Tal, L. El Ghaoui, A. Nemirovski, Robust optimization, Princeton University Press, 2009.

[51] V. Gabrel, C. Murat, A. Thiele, Recent advances in robust optimization: An overview, European Journal of Operational Research 235 (3) (2014) 471–483.

[52] A. Ben-Tal, A. Nemirovski, Robust optimization–methodology and applications, Mathematical Programming 92 (3) (2002) 453–480.

[53] H.-G. Beyer, B. Sendhoff, Robust optimization–a comprehensive survey, Computer methods in applied mechanics and engineering 196 (33) (2007) 3190–3218.

[54] J. M. Mulvey, R. J. Vanderbei, S. A. Zenios, Robust optimization of large-scale systems, Operations research 43 (2) (1995) 264–281.

[55] E. Erdoğan, G. Iyengar, Ambiguous chance constrained problems and robust optimization, Mathematical Programming 107 (1-2) (2006) 37–61.

[56] L. E. Ghaoui, M. Oks, F. Oustry, Worst-case value-at-risk and robust portfolio optimization: A conic programming approach, Operations Research 51 (4) (2003) 543–556.

[57] M. Chu, Y. Zinchenko, S. G. Henderson, M. B. Sharpe, Robust optimization for intensity modulated radiation therapy treatment planning under uncertainty, Physics in Medicine and Biology 50 (23) (2005) 5463.

[58] T. Bortfeld, T. C. Chan, A. Trofimov, J. N. Tsitsiklis, Robust management of motion uncertainty in intensity-modulated radiation therapy, Operations Research 56 (6) (2008) 1461–1473.

[59] F. Meng, J. Qi, M. Zhang, J. Ang, S. Chu, M. Sim, A robust optimization model for managing elective admission in a public hospital, Operations Research 63 (6) (2015) 1452–1467.

[60] A. Sauré, M. A. B. J. Patrick, Dynamic multi-priority, multi-class patient scheduling with stochastic service times, European Journal of Operational Research 280 (1) (2020) 254–265.

[61] C. Zacharias, N. Liu, M. A. Begen, Dynamic inter-day and intra-day scheduling, Available at SSRN 3728077 (2020).

[62] S. Mittal, A. S. Schulz, S. Stiller, Robust Appointment Scheduling, in: K. Jansen, J. D. P. Rolim, N. R. Devanur, C. Moore (Eds.), Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014), Vol. 28 of Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2014, pp. 356–370.

[63] J. Qi, Mitigating delays and unfairness in appointment systems, Management Science 63 (2) (2017) 566–583.

[64] R. Jiang, S. Shen, Y. Zhang, Integer programming approaches for appointment scheduling with random no-shows and service durations, Operations Research 65 (6) (2017) 1638–1656.

[65] K. S. Shehadeh, A. E. Cohn, R. Jiang, A distributionally robust optimization approach for outpatient colonoscopy scheduling, European Journal of Operational Research 283 (2) (2020) 549–561.

[66] D. Strum, J. May, L. Vargas, Modeling the uncertainty of surgical procedure times: Comparison of log-normal and normal models, Anesthesiology 92 (4) (2000) 1160–1167.

[67] R. C. Dahiya, I. Guttman, Shortest confidence and prediction intervals for the log-normal, The Canadian Journal of Statistics 10 (1982) 277–291.

[68] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2019).

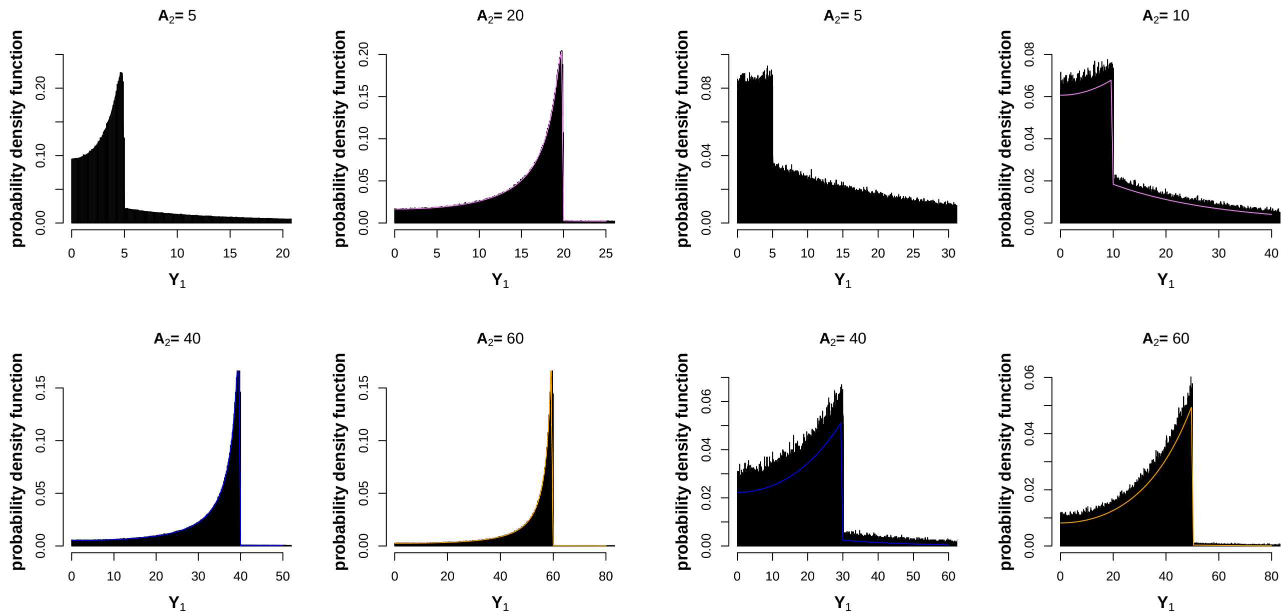
[69] R. P. Brent, Algorithms for minimization without derivatives, Prentice-Hall, Englewood Cliffs, New Jersey, 2013.

[70] R. J. Tibshirani, Fast computation of the median by successive binning, Unpublished manuscript, http://stat. stanford. edu/ryantibs/median (2008).

[71] R. Barzanji, B. Naderi, M. A. Begen, Decomposition algorithms for the integrated process planning and scheduling problem, Omega 93 (2020) 102025.

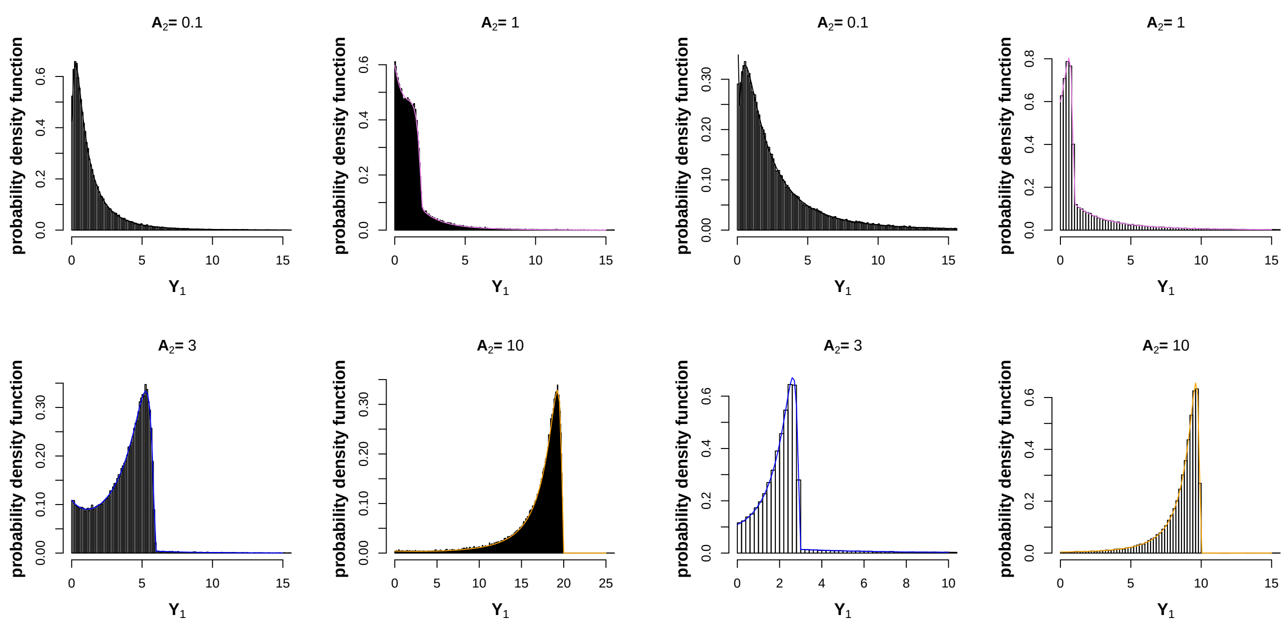
[72] B. C. Gencosman, M. A. Begen, H. C. Ozmutlu, I. O. Yilmaz, Scheduling methods for efficient stamping operations at an automotive company, Production and Operations Management 25 (11) (2016) 1902–1918.

## Appendix




(a) (b)

Figure 7: (a) Probability density function of the total cost for different values of $A_2$. For all four panels, we set the parameters as $\alpha = 1,\ \mu = 1.58,\ \sigma = 1.68$ in the lognormal distribution such that its mean and standard deviation are 20 and 5, respectively. (b) Probability density function of the total cost for different values of $A_2$ with an exponentially distributed service time of mean 20.




(a) (b)

Figure 8: (a) Probability density function of the total cost for different values of $A_2$. For all four panels, we set the parameters as $o_1 = 1,\ u_1 = 2,\ \mu = 0\ \sigma = 1$ in the lognormal distribution. (b) The same set up as in panel (a) except that $o_1 = 2, u_1 = 1$.

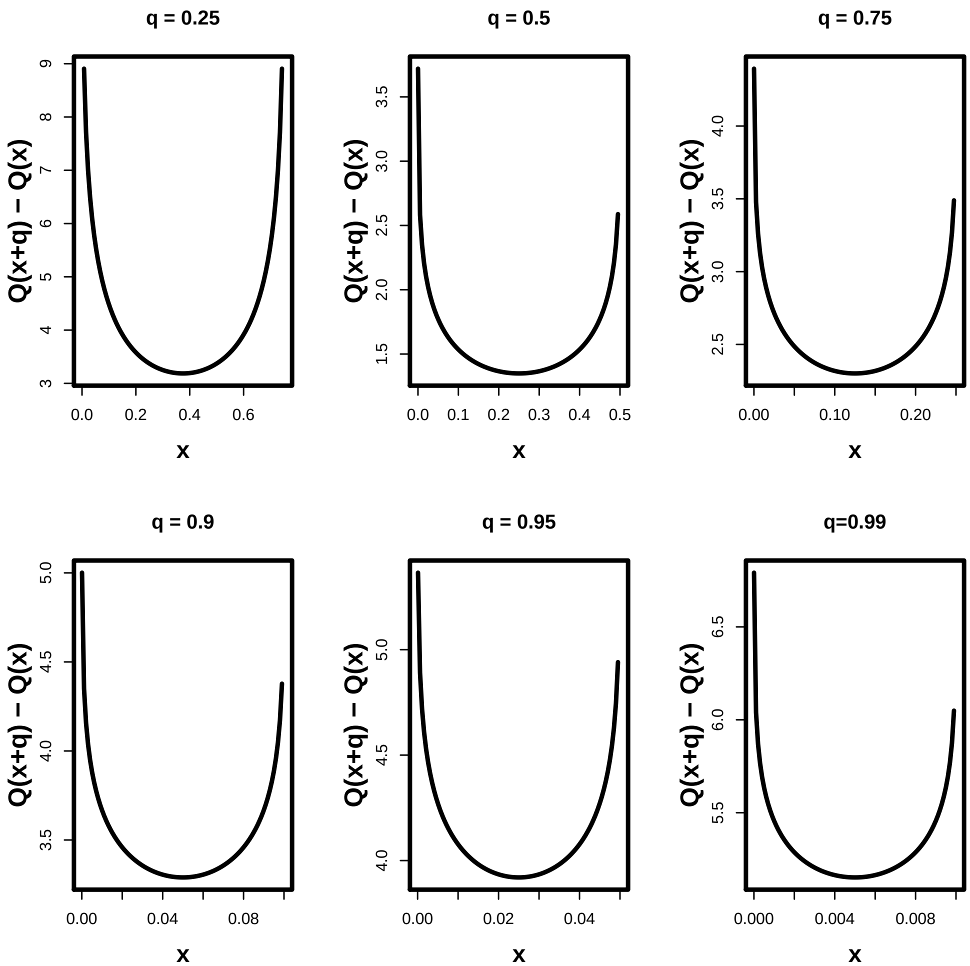



Figure 9: $Q_{U_1}(x+q) - Q_{U_1}(x)$ changes with $x$ when the value of $q$ varies. The service duration is assumed to follow Normal(20, 25).

**Newton-type algorithm**: We aim to find the minimizer of $Q_{U_1}(x+q) - Q_{U_1}(x)$ as a function of $x$. Here $U_1$ is the random variable used to denote the duration time of a job, which is assumed to follow a lognormal distribution in our work. Thus $Q_{U_1}$ is the quantile function of a lognormal distribution, i.e., the inverse function of the cumulative distribution of a lognormal distribution.

Let $f(x) = Q_{U_1}(x+q) - Q_{U_1}(x)$. Thus the first derivative of $f$ is $f'(x) = Q'_{U_1}(x+q) - Q'_{U_1}(x)$, where $Q'_{U_1}(x)$ is the first derivative of $Q_{U_1}(x)$. The Newton-type algorithm works in the following way:

**Algorithm 1 : Newton-type Algorithm to Find the Minimizer of** $Q_{U_1}(x+q) - Q_{U_1}(x)$

*Step* 1: Take an initial value $x_0$ from $(0, 1-q)$. Then $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$

*Step* 2: For $k = 1, 2, \ldots, K$, $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$, where $K$ denotes the largest number of iterations.

We will stop iterations in Step 2 if we find in several consecutive steps there is little change in $x_k$. Such $x_{k^\star}$ is regarded as the limit of this sequence and the minimizer of the function $f$. We use the R function **nlm** ([68]) to implement this algorithm.

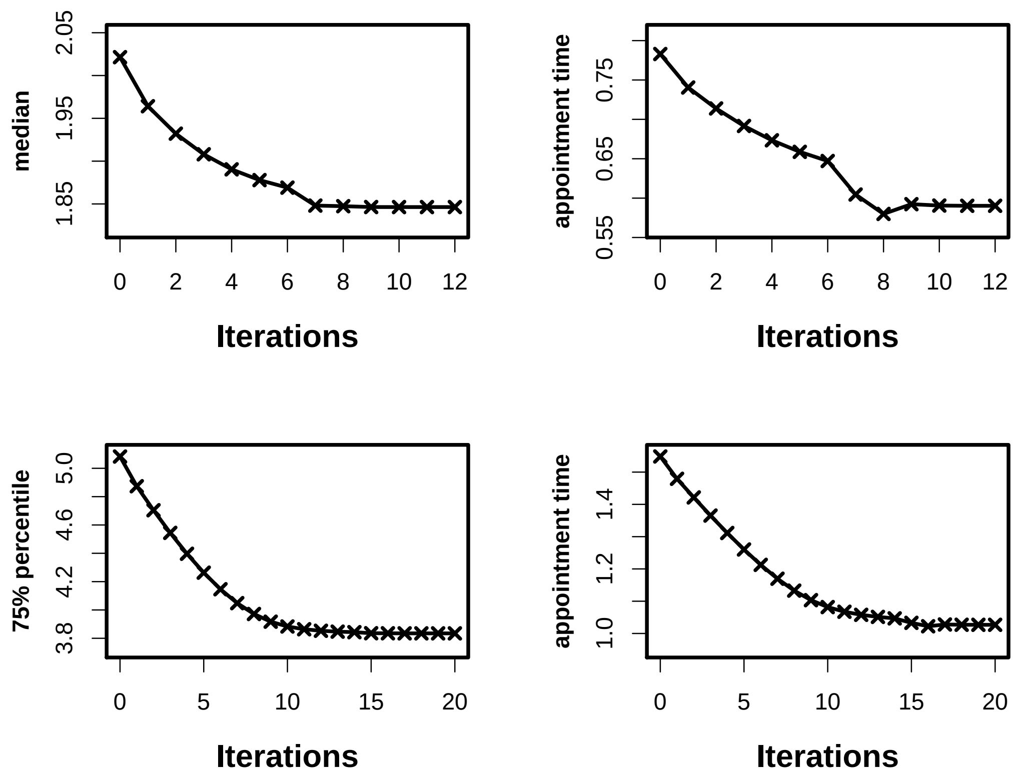



Figure 10: The trace of the algorithm over iterations. The two left panels show the median and the 75% percentile of the total cost $Y_1$ defined in (1) in each iteration of the algorithm. The two right panels show how the appointment time changes as the algorithm iterates when the objective is to minimize the median and 75% percentile of the total cost. We set the constant parameters as $\alpha = 2$, $\mu = 0$, and $\sigma = 1$.

The upper left panel in Figure 10 shows the median of the total cost $Y_1$ defined in (1) in each iteration of the algorithm when $q = 0.5$, $\alpha = 2$, $\mu = 0$ and $\sigma = 1$. The median of the total cost keeps decreasing as the algorithm proceeds and converges in the end. The upper right panel in the figure shows how the appointment time changes as the algorithm iterates. It goes down to the bottom at first but then goes up. In the end, it becomes stable and converges. Similar patterns are observed when minimizing the 0.75th quantile of the total cost, which is shown in the lower two panels of Figure 10.

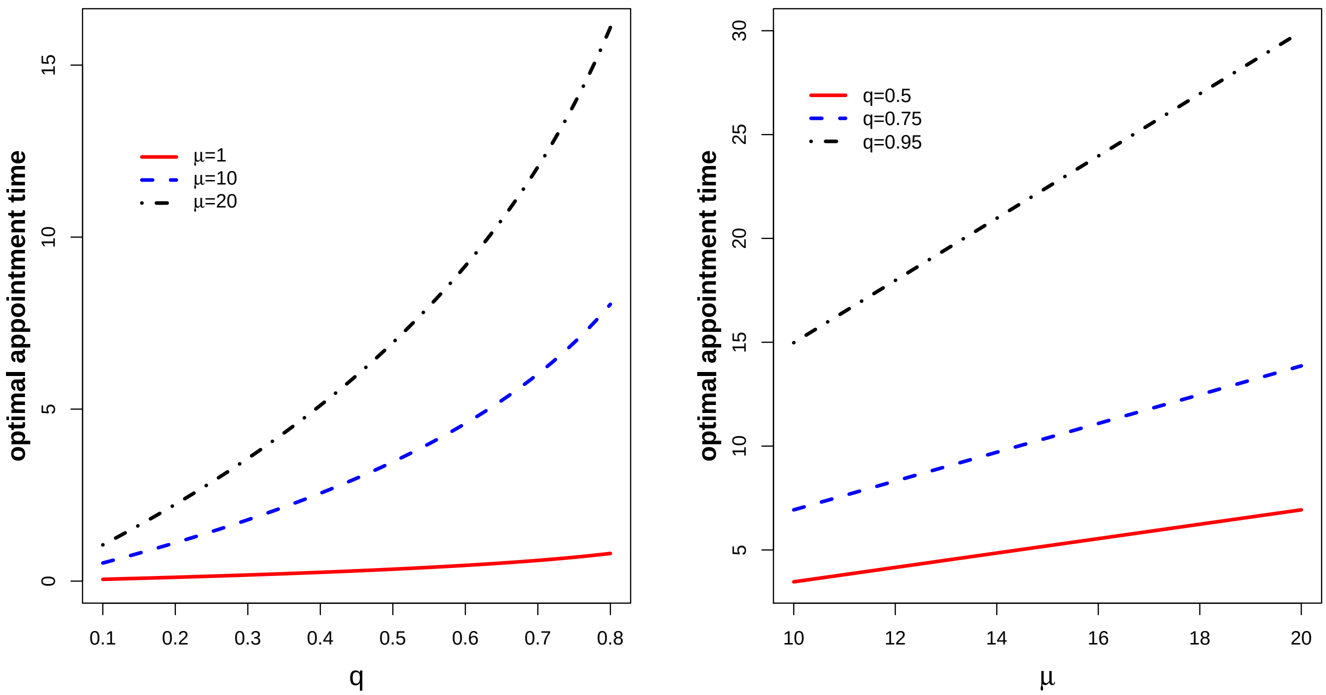



Figure 11: The optimal appointment time under different $\mu$ and $q$ values. For both left and right panels, we set the constant parameters as $\alpha = 2$ and the duration time follows an exponential distribution with $\mu$ being the mean.

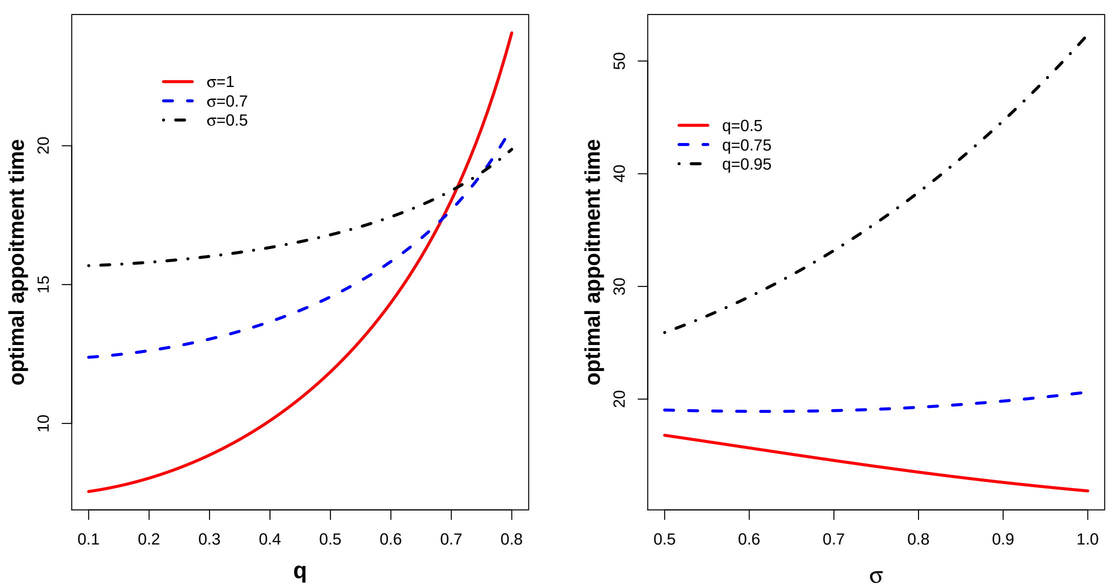



Figure 12: The optimal appointment time under different $\sigma$ and $q$ values. For both left and right panels, we set the constant parameters as $\alpha = 2$ and $\mu = 3$.

| | $n = 100$ | $n = 1000$ | $n = 10000$ | *Real* |
|---|---|---|---|---|
| $q = 0.2$ | 0.89 | 1.04 | 0.68 | 0.64 |
| $q = 0.5$ | 2.93 | 3.09 | 2.49 | 2.44 |
| $q = 0.8$ | 8.12 | 10.12 | 10.06 | 10.01 |

Table 13: Comparison of the sample-based optimal appointment time and the real optimal appointment time when the number of samples $n = 100, 1,000, 10,000$. The optimization objective is to minimize the $q$th quantiles of the total cost. The real optimal appointment time is acquired from the algorithm introduced in Section 2.1, while assuming the underlying distribution of the service duration is the log-normal distribution. We set the constant parameters as $\alpha = 2$, $\mu = 1.58$ and $\sigma = 1.68$ such that the mean and the standard deviation of the log-normal distribution is 20 and 5, respectively.

| | $n = 100$ | $n = 1000$ | $n = 10000$ | *Real* |
|---|---|---|---|---|
| $q = 0.2$ | 2.08 | 2.42 | 2.28 | 2.23 |
| $q = 0.5$ | 7.84 | 7.11 | 7.01 | 6.93 |
| $q = 0.8$ | 15.44 | 16.22 | 16.18 | 16.09 |

Table 14: Comparison of the sample-based optimal appointment time and the real optimal appointment time when the number of samples $n = 100, 1,000, 10,000$. The optimization objective is to minimize the $q$th quantiles of the total cost. The real optimal appointment time is acquired from the algorithm introduced in Section 2.1, while assuming the underlying distribution of the service duration is an exponential distribution. We set the constant parameters as $\alpha = 2$, $\lambda = 1/20$ such that the mean of the exponential distribution is 20.

| | $(o_1, u_1) = (1, 2)$ | | $(o_1, u_1) = (2, 1)$ | | $(o_1, u_1) = (1, 3)$ | | $(o_1, u_1) = (3, 1)$ | |
|---|---|---|---|---|---|---|---|---|
| | T | S | T | S | T | S | T | S |
| $q = 0.2$ | 0.45 | 0.46 | 0.83 | 0.84 | 0.35 | 0.37 | 0.99 | 0.98 |
| $q = 0.5$ | 1.63 | 1.65 | 3.25 | 3.28 | 1.23 | 1.24 | 3.89 | 3.93 |
| $q = 0.8$ | 6.67 | 6.62 | 13.35 | 13.23 | 5.01 | 4.96 | 16.01 | 15.87 |

Table 15: The optimal appointment time for $q$th quantiles of the total cost with unequal cost coefficients. The total cost $Y_1 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+$. "T" stands for optimization based on the theoretical distribution of the total cost; while "S" stands for optimization based on the empirical distribution of the total cost.

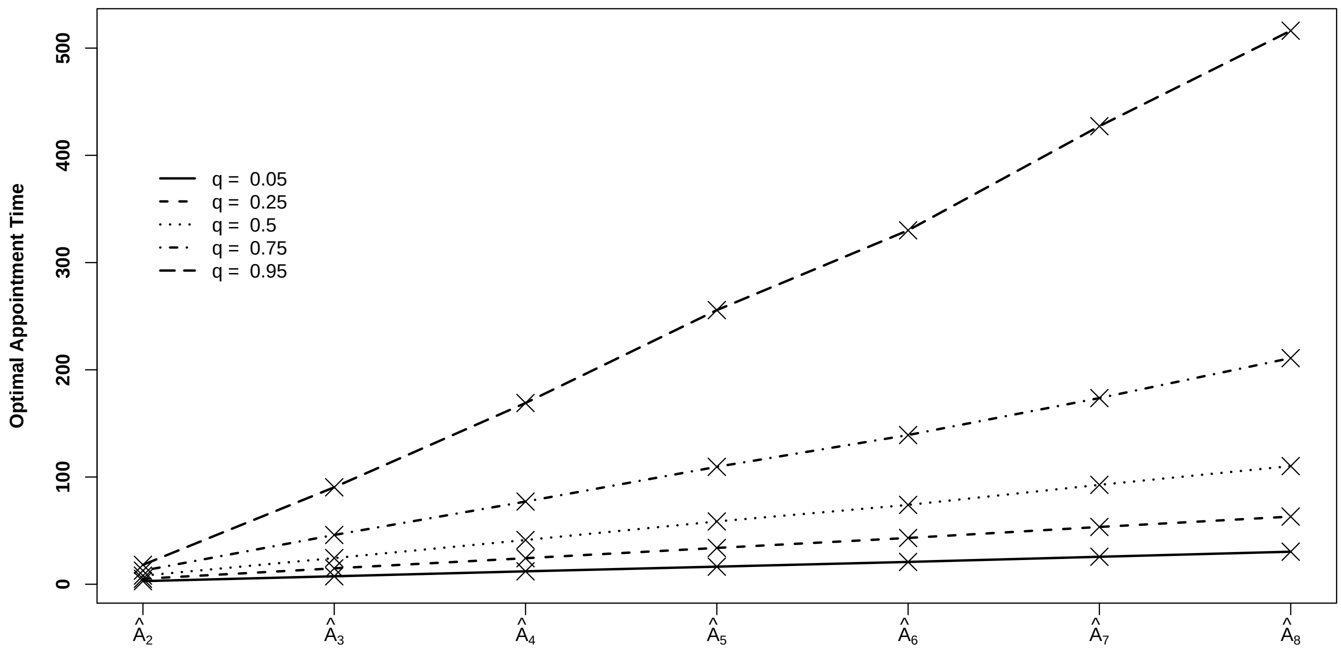
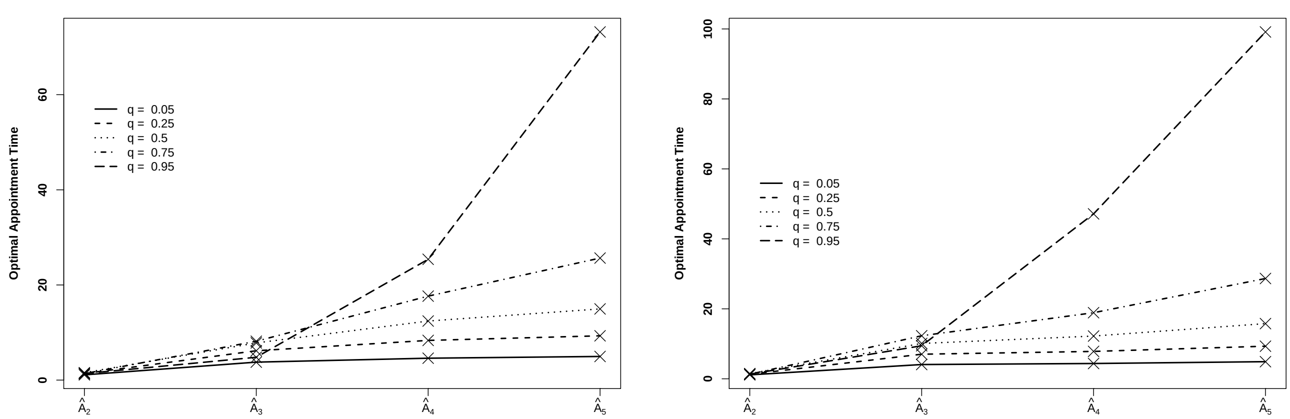



Figure 13: The optimal appointment time of seven jobs under different quantile levels ($q$). Duration times are independently sampled from logNormal with mean = 20 and standard deviation = 5 and all cost coefficients are set as 1.




(a)

(b)

Figure 14: (a) The optimal appointment times with respect to the first (and second) sequencing rule. (b) The optimal appointment times with respect to the third sequencing rule.

| $(\alpha, \beta)$ | $q = 0.05$ | $q = 0.25$ | $q = 0.50$ | $q = 0.75$ | $q = 0.95$ |
|---|---|---|---|---|---|
| (9, 1) | 56.0 | 221.3 | 633.5 | 1052.8 | 1968.8 |
| (8, 2) | 72.2 | 241.7 | 654.6 | 1089.5 | 2039.2 |
| (7, 3) | 83.3 | 256.3 | 654.6 | 1101.0 | 2116.4 |
| (6, 4) | 87.0 | 264.3 | 649.8 | 1098.3 | 2155.8 |
| (5, 5) | 87.5 | 271.7 | 645.8 | 1087.5 | 2258.6 |
| (4, 6) | 86.0 | 284.1 | 626.1 | 1069.1 | 2359.4 |
| (3, 7) | 82.2 | 294.4 | 596.5 | 1027.5 | 2496.5 |
| (2, 8) | 74.0 | 303.0 | 546.3 | 962.7 | 2628.9 |
| (1, 9) | 59.5 | 244.0 | 464.4 | 851.7 | 2793.4 |

Table 16: Summary of $q$-th quantiles of the total cost, corresponding to the two-job optimal appointment time shown in Table 8. The total cost $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$.

| $(o_1, u_1)$ | $q = 0.05$ | | $q = 0.25$ | | $q = 0.50$ | | $q = 0.75$ | | $q = 0.95$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ |
| (9, 1) | 60.1 | 113.2 | 80.6 | 171.3 | 179.9 | 290.8 | 264.6 | 419.0 | 415.8 | 956.1 |
| (8, 2) | 59.6 | 111.1 | 67.8 | 146.6 | 158.0 | 259.6 | 222.4 | 368.4 | 347.6 | 819.6 |
| (7, 3) | 59.6 | 109.4 | 57.1 | 132.1 | 134.0 | 233.9 | 176.5 | 326.4 | 257.1 | 672.9 |
| (6, 4) | 58.6 | 106.4 | 53.0 | 125.0 | 114.9 | 211.9 | 163.5 | 289.5 | 167.4 | 545.4 |
| (5, 5) | 57.0 | 103.0 | 46.3 | 117.8 | 102.3 | 191.8 | 150.7 | 253.2 | 149.3 | 470.8 |
| (4, 6) | 57.4 | 98.4 | 41.4 | 107.4 | 83.0 | 156.4 | 122.9 | 215.5 | 123.3 | 383.1 |
| (3, 7) | 52.6 | 95.9 | 32.3 | 91.8 | 75.0 | 145.6 | 95.9 | 175.8 | 103.0 | 297.9 |
| (2, 8) | 49.3 | 94.6 | 25.3 | 78.5 | 50.0 | 106.4 | 69.6 | 133.7 | 80.2 | 212.2 |
| (1, 9) | 46.0 | 90.2 | 15.9 | 71.3 | 26.6 | 76.6 | 39.6 | 88.3 | 51.4 | 119.3 |

Table 17: Summary of the two-job optimal appointment times with various combinations of cost coefficients and probability levels ($q$). The total cost $Y_2 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ + o_2(C_2 - A_3)^+ + u_2(A_3 - C_2)^+$. Here we assume that $o_1 = o_2$ and $u_1 = u_2$.

| $(o_1, u_1)$ | $q = 0.05$ | $q = 0.25$ | $q = 0.50$ | $q = 0.75$ | $q = 0.95$ |
|---|---|---|---|---|---|
| (9, 1) | 33.2 | 110.3 | 228.8 | 383.0 | 936.1 |
| (8, 2) | 59.9 | 191.1 | 397.2 | 659.0 | 1599.2 |
| (7, 3) | 79.1 | 243.4 | 518.7 | 862.1 | 1959.0 |
| (6, 4) | 90.0 | 276.0 | 607.5 | 1002.0 | 2101.7 |
| (5, 5) | 95.0 | 289.2 | 654.2 | 1070.8 | 2101.7 |
| (4, 6) | 92.7 | 284.7 | 650.4 | 1058.7 | 2154.7 |
| (3, 7) | 82.5 | 257.7 | 592.2 | 957.8 | 1917.5 |
| (2, 8) | 64.1 | 204.3 | 467.2 | 757.9 | 1489.4 |
| (1, 9) | 36.8 | 120.8 | 275.9 | 444.0 | 857.4 |

Table 18: Summary of $q$-th quantiles of the total cost, corresponding to the two-job optimal appointment time shown in Table 17. The total cost $Y_2 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ + o_2(C_2 - A_3)^+ + u_2(A_3 - C_2)^+$.

| $(o_1, u_1)$ | Median | | | Mean | | |
|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $Q_{Y_2}$ | $\hat{A}_2$ | $\hat{A}_3$ | mean |
| (9, 1) | 179.9 | 290.8 | 228.8 | 285.0 | 646.0 | 729.9 |
| (8, 2) | 158.0 | 259.6 | 397.2 | 236.0 | 456.0 | 973.7 |
| (7, 3) | 134.0 | 233.9 | 518.7 | 207.0 | 353.0 | 1074.8 |
| (6, 4) | 114.9 | 211.9 | 607.5 | 183.0 | 302.0 | 1098.5 |
| (5, 5) | 102.3 | 191.8 | 654.2 | 154.0 | 258.0 | 1067.2 |
| (4, 6) | 83.0 | 156.4 | 650.4 | 123.0 | 216.0 | 982.7 |
| (3, 7) | 75.0 | 145.6 | 592.2 | 81.0 | 166.0 | 832.8 |
| (2, 8) | 50.0 | 106.4 | 467.2 | 55.0 | 121.0 | 613.1 |
| (1, 9) | 26.6 | 76.6 | 275.9 | 31.0 | 85.0 | 338.5 |

Table 19: Two-job optimal appointment times with respect to median and expected cost objectives with different combinations of cost coefficients. The total cost $Y_2 = o_1(C_1 - A_2)^+ + u_1(A_2 - C_1)^+ + o_2(C_2 - A_3)^+ + u_2(A_3 - C_2)^+$. $Q_{Y_2}$ denotes the corresponding median of the total cost under the optimal appointment time.

| $(\alpha, \beta)$ | $q = 0.05$ | | $q = 0.25$ | | $q = 0.50$ | | $q = 0.75$ | | $q = 0.95$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ | $\hat{A}_2$ | $\hat{A}_3$ |
| (9, 1) | 12.0 | 23.5 | 11.0 | 23.5 | 11.0 | 24.0 | 12.0 | 27.0 | 16.1 | 35.1 |
| (8, 2) | 12.0 | 23.5 | 10.6 | 23.5 | 10.8 | 23.8 | 12.0 | 28.0 | 15.1 | 39.5 |
| (7, 3) | 11.9 | 22.9 | 11.0 | 23.7 | 10.8 | 24.7 | 11.9 | 27.3 | 14.4 | 39.9 |
| (6, 4) | 11.0 | 21.8 | 12.0 | 24.3 | 11.3 | 24.3 | 11.8 | 27.5 | 15.1 | 39.4 |
| (5, 5) | 11.7 | 24.0 | 11.7 | 23.7 | 11.3 | 24.3 | 12.0 | 27.5 | 16.3 | 37.8 |
| (4, 6) | 11.0 | 23.5 | 12.0 | 24.2 | 11.3 | 24.0 | 13.0 | 27.2 | 18.4 | 39.1 |
| (3, 7) | 12.0 | 24.1 | 11.0 | 23.2 | 12.6 | 25.8 | 14.3 | 27.2 | 17.2 | 37.6 |
| (2, 8) | 12.3 | 24.0 | 11.8 | 23.8 | 14.3 | 27.0 | 16.7 | 29.0 | 21.8 | 41.3 |
| (1, 9) | 10.1 | 23.0 | 11.0 | 23.5 | 14.6 | 26.6 | 17.7 | 30.1 | 25.1 | 43.4 |

Table 20: Summary of the optimal appointment time for two identical jobs with various combinations of cost coefficients and probability levels. The total cost $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$.

| $(\alpha, \beta)$ | $q = 0.05$ | $q = 0.25$ | $q = 0.50$ | $q = 0.75$ | $q = 0.95$ |
|---|---|---|---|---|---|
| (9, 1) | 2.5 | 14.0 | 29.0 | 49.5 | 112.2 |
| (8, 2) | 5.0 | 16.6 | 30.2 | 50.0 | 116.5 |
| (7, 3) | 6.6 | 18.0 | 31.3 | 50.8 | 121.2 |
| (6, 4) | 7.0 | 19.0 | 31.5 | 51.5 | 125.8 |
| (5, 5) | 6.7 | 18.3 | 31.7 | 52.5 | 129.2 |
| (4, 6) | 7.0 | 19.1 | 33.1 | 55.1 | 134.1 |
| (3, 7) | 6.5 | 18.5 | 33.9 | 57.4 | 138.7 |
| (2, 8) | 5.7 | 17.7 | 33.7 | 58.3 | 139.7 |
| (1, 9) | 2.9 | 16.5 | 33.1 | 58.2 | 140.3 |

Table 21: Summary of $q$th quantiles of the total cost, corresponding to the two-job appointment time in Table 20. The total cost $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$.

| $(\alpha, \beta)$ | Median | | | Mean | | |
|---|---|---|---|---|---|---|
| | $\hat{A}_2$ | $\hat{A}_3$ | $Q_{Y_2}$ | $\hat{A}_2$ | $\hat{A}_3$ | mean |
| (9, 1) | 11.0 | 24.0 | 29.0 | 12.0 | 27.0 | 49.1 |
| (8, 2) | 10.8 | 23.8 | 30.2 | 12.0 | 27.0 | 51.4 |
| (7, 3) | 10.8 | 24.7 | 31.3 | 13.0 | 27.0 | 53.5 |
| (6, 4) | 11.3 | 24.3 | 31.5 | 13.0 | 27.0 | 55.5 |
| (5, 5) | 11.3 | 24.3 | 31.7 | 13.0 | 27.0 | 57.5 |
| (4, 6) | 11.3 | 24.0 | 33.1 | 14.0 | 28.0 | 59.5 |
| (3, 7) | 12.6 | 25.8 | 33.9 | 14.0 | 28.0 | 61.2 |
| (2, 8) | 14.3 | 27.0 | 33.7 | 16.0 | 29.0 | 62.4 |
| (1, 9) | 14.6 | 26.6 | 33.1 | 19.0 | 32.0 | 62.3 |

Table 22: Summary of the identical two-job optimal appointment times for median and expected cost optimization. The total cost $Y_2 = \alpha|C_1 - A_2| + \beta|C_2 - A_3|$, and $Q_{Y_2}$ denotes the median or the total cost under the optimal appointment times.