Appointment Scheduling with Restricted People

1 Preliminary Study

The service time for customer i, ξ_i, stochastic with a mean of μ_i and a standard deviation of σ_i. The service times are mutually independent. For each customer $i = 1, \ldots, n$, we use A_i to denote the appointment time, $S_i = \max\{A_i, S_{i-1} + \xi_{i-1}\}$ denote the actual starting time of service. We assume that the customers will arrive at the appointed time. Especially, $A_1 = S_1 = 0$.

The waiting time for customer i is $S_i - A_i$, the total waiting time is $\sum_{i=2}^{n} \alpha_i (S_i - A_i)$, where α_i is the weight for customer i. The overtime is $(S_n + \xi_n - T)^+$ and the total idle time is $\sum_{i=1}^{n-1}[S_{i+1} - (S_i + \xi_i)] = S_n - \sum_{i=1}^{n-1} \xi_i$.

In the scenario with at least 2 customers overlapping in the waiting room, we can calculate the overlapping time. Let t_{ij} denote the overlapping time between two customers i and j. Then, $t_{i,j} = (S_i - A_j)^+$, indicating there are at least $(j - i + 1)$ customers waiting.

The duration when there are only $(j - i + 1)$ people from customer i to customer j are waiting is $t_{i,j} - t_{i,j+1}$, $i = 2, \ldots, n-1, j \geq i$.

Total overlapping time: $\sum_{i=2}^{n-1} \sum_{j=i}^{n-1} \gamma_{i,j}(t_{i,j} - t_{i,j+1})$

Problem to minimize the total time cost:

$$\begin{aligned} \min_{\mathbf{A}} \quad & E_\xi \left[\left(S_n - \sum_{i=1}^{n-1} \xi_i \right) + \sum_{i=2}^{n-1} \sum_{j=i}^{n-1} \gamma_{i,j}(t_{i,j} - t_{i,j+1}) + \beta(S_n + \xi_n - T)^+ \right] \\ \text{s.t.} \quad & S_i = \max\{A_i, S_{i-1} + \xi_{i-1}\} \\ & S_1 = 0 \end{aligned} \tag{1}$$

To minimize the makespan:

$$\begin{aligned} \min_{\mathbf{A}} \quad & E_\xi \left(S_n + \xi_n \right) \\ \text{s.t.} \quad & E_\xi \left(t_{i,j} - t_{i,j+1} \right) \leq L_{ij}, i = 2, \ldots, n-1, j \geq i \end{aligned} \tag{2}$$

L_{ij} indicates constraint on the duration of $(j - i + 1)$ people waiting.

2 Model

We redefine the scheduling problem using the following notation. Let $\mathbf{\Delta} = (\Delta_1, \ldots, \Delta_n)$ denote the **appointment intervals**, where Δ_i is the time allocated between the start of customer i and customer

$i+1$. Let $\mathbf{A} = (A_1, \ldots, A_n)$ represent the **scheduled appointment times**, with $A_i = \sum_{k=1}^{i-1} \Delta_k$ (assuming $A_1 = 0$). Let $\mathbf{Z} = (Z_1, \ldots, Z_n)$ be the **random service durations**, where Z_i is the stochastic service time for customer i.

The waiting time for customer i is recursively defined as:

$$\begin{aligned} W_i(\mathbf{Z}_{i-1}, \mathbf{\Delta}_{i-1}) &= [A_{i-1} + W_{i-1}(\mathbf{Z}_{i-2}, \mathbf{\Delta}_{i-2}) + Z_{i-1} - A_i]^+ \\ &= [W_{i-1}(\mathbf{Z}_{i-2}, \mathbf{\Delta}_{i-2}) + Z_{i-1} - \Delta_{i-1}]^+, \end{aligned}$$

where $[*]^+ = \max(*, 0)$. $W_1(\mathbf{Z}_0, \mathbf{\Delta}_0) = 0$.

Let W_{ij} denote the **simultaneous waiting duration** for customers i through j, meaning the length of time during which all customers from i to j are simultaneously waiting. This can be expressed recursively as: $W_{i,j}(\mathbf{Z}_{i-1}, \mathbf{\Delta}_{j-1}) = [W_{i-1}(\mathbf{Z}_{i-2}, \mathbf{\Delta}_{i-2}) + Z_{i-1} - \sum_{k=i-1}^{j-1} \Delta_k]^+$, where $W_i \equiv W_{i,i}$ is the individual waiting time for customer i. W_{ij} captures the time window where customers i to j all experience waiting simultaneously due to delays from earlier customers (1 to $i-1$) and insufficient buffer times.

The finish time for customer i is $T_i(\mathbf{Z}_i, \mathbf{\Delta}_{i-1}) = A_i + W_i(\mathbf{Z}_{i-1}, \mathbf{\Delta}_{i-1}) + Z_i$. We aim to minimize the total schedule span, i.e., T_n, subject to constraints on individual and group waiting times.

The formulation of the problem can be expressed as follows

$$\begin{aligned} \min \quad & E\left[T_n(\mathbf{Z}_n, \mathbf{\Delta}_{n-1})\right] \\ \text{s.t.} \quad & E\left[W_{i,j}(\mathbf{Z}_{i-1}, \mathbf{\Delta}_{j-1})\right] \le w_{ij}, i = 2, \ldots, n-1, j \ge i \end{aligned} \tag{3}$$

In this setting, w_{ij} is related with the number of customers from i to j, i.e., $j - i + 1$. We can use w_k to indicate the upper limit on the time when there are k customers waiting.

$T_n(\mathbf{Z}_n, \mathbf{\Delta}_{n-1}) = A_n + W_n(\mathbf{Z}_{n-1}, \mathbf{\Delta}_{n-1}) + Z_n$.

Lemma 1. *For any given realization of $\mathbf{Z}_n$, $T_n(\mathbf{Z}_n, \mathbf{\Delta}_{n-1})$ becomes shorter when some customer is scheduled to arrive earlier while the schedule for others remain unchanged.*

The optimal schedule can be obtained by minimizing Δ_i.

When $i = 1$, the first customer doesn't need to wait, i.e., $W_1(\mathbf{Z}_0, \mathbf{\Delta}_0) = 0$.

When $i = 2$, only one constraint $E\left[W_1(\mathbf{Z}_0, \mathbf{\Delta}_0) + Z_1 - \Delta_1\right]^+ \le w_1$ is applied, then Δ_1^* can be obtained.

When $i = 3$, there are two constraints on the waiting time of the third customer.

$E\left[W_2(\mathbf{Z}_1, \mathbf{\Delta}_1^*) + Z_2 - \Delta_2\right]^+ \le w_1$.

$E\left[W_1(\mathbf{Z}_0, \mathbf{\Delta}_0) + Z_1 - \Delta_1^* - \Delta_2\right]^+ \le w_2$.

Then Δ_2^* can be obtained.

When $i = 4$, there are two constraints on the waiting time.

$E\left[W_3(\mathbf{Z}_2, \mathbf{\Delta}_2^*) + Z_3 - \Delta_3\right]^+ \le w_1$.

$E\left[W_2(\mathbf{Z}_1, \mathbf{\Delta}_1) + Z_2 - \Delta_2^* - \Delta_3\right]^+ \le w_2$.

Proposition 1. *By solving the above problems sequentially, the optimal schedule can be obtained.*

Then we analyze these problems. The function on the left-hand side is decreasing in the variable Δ_i.

When $\Delta_1 = 0$, $E\left[W_1(\mathbf{Z}_0, \mathbf{\Delta}_0) + Z_1 - \Delta_1\right]^+ = E\left[Z_1\right]^+$. If $E\left[Z_1\right]^+ \leq w_1$, $\Delta_1^* = 0$; if $E\left[Z_1\right]^+ > w_1$, $E\left[Z_1 - \Delta_1^*\right]^+ = w_1$.

If Z_i follows from the exponential distribution with rate λ, $E\left[Z_1 - \Delta_1\right]^+ = \frac{1}{\lambda} e^{-\lambda \Delta_1}$, then

$$\Delta_1^* = \begin{cases} -\frac{\ln(\lambda w_1)}{\lambda}, & \text{if } \lambda w_1 < 1 \\ 0, & \text{if } \lambda w_1 \geq 1 \end{cases}$$

3 Literature

1. Possible traits: heterogeneous customers, no-show, lateness, walk-in

Different models: objective: minimize the total cost, minimize the makespan (the departure time of the last customer).

Traditional Appointment Scheduling Model.

1. with overbooking and no-shows (partial punctuality)

- discrete n time slots.

- minimize the waiting cost, idle time and overtime costs.

- analyze three components seperately

2. Under a service-level constraint (waiting time threshold)

- makespan

- the optimal schedule can be obtained sequentially.

4 Some Instances

We evaluate and compare three constraint scenarios:

- Waiting Time Constraint: Limits the maximum waiting time for each customer.
- Overlapping Time Constraint: Restricts two simultaneous waiting overlaps.
- Combined Constraints: Enforces both waiting time and overlapping time limits.

In all cases, the schedule times of later customers do not affect the appointments of preceding customers, ensuring feasibility in sequential scheduling.

Setting the mean of simultaneous waiting time to be no larger than zero (requiring that the number of people never exceeds a certain threshold) is an overly stringent condition. In systems without a maximum service time constraint, this requirement would significantly delay the scheduled time compared to systems that only consider the single waiting time. Moreover, when extreme situations occur (the service times of the earlier customers are extremely large), the scheduled time would be postponed even further.

An alternative approach is to set a threshold on the probability of the number of waiting customers not exceeding a certain limit.

The original model considers the overlapping time,

$$\begin{aligned} \min \quad & E[T_n(\mathbf{Z}_n, \mathbf{\Delta}_{n-1})] \\ \text{s.t.} \quad & E[W_{i,i+1}(\mathbf{Z}_{i-1}, \mathbf{\Delta}_i)] \leq w, i = 2, \ldots, n-1 \end{aligned} \tag{4}$$

Wells-Riley model: The probability P of infection in a shared space is given by:

$P = 1 - e^{-I \cdot q \cdot p \cdot t/Q}$

I: Number of infectious individuals q: Quanta emission rate p: Mask penetration factor t: Exposure time Q: Room ventilation rate

The overlapping constraint is not stringent/sufficient. Although setting $\Delta_1 = 0$ minimizes T_n, increasing Δ_1 can significantly reduce Customer 2's waiting time while only slightly increasing T_n.

Table 1: Schedule Intervals and Times

Waiting Model	$\Delta_1 = 0.72$	$\Delta_2 = 31.1$	$\Delta_3 = 31.3$	$\Delta_4 = 32.3$	$\Delta_5 = 32.5$	$T_n = 188.76$
Waiting	0	30	30	30	30	30
Overlapping	0	5.2	7.2	8.2	8.9	
Overlapping Model	$\Delta_1 = 0.1$	$\Delta_2 = 27.4$	$\Delta_3 = 35.1$	$\Delta_4 = 34.7$	$\Delta_5 = 34.6$	$T_n = 189.53$
Waiting	0	30.6	34.0	30.3	30.5	26.8
Overlapping	0	7.4	7.4	7.4	7.4	
Waiting Model	$\Delta_1 = 0.72$	$\Delta_2 = 31.1$	$\Delta_3 = 31.3$	$\Delta_4 = 32.3$	$T_n = 156.36$	
Waiting	0	30	30	30	30	total(120)
Overlapping	0	5.2	7.2	8.2		total(20.6)
Overlapping Model	$\Delta_1 = 0.42$	$\Delta_2 = 27.11$	$\Delta_3 = 35.1$	$\Delta_4 = 34.7$	$T_n = 156.56$	
Waiting	0	30.5	34.0	32.1	28.3	total(125.0)
Overlapping	0	7.4	7.4	7.4		total(22.2)
Overlapping (fixed Δ_1)	$\Delta_1 = 5$	$\Delta_2 = 22.53$	$\Delta_3 = 35.12$	$\Delta_4 = 34.67$	$T_n = 156.58$	
Waiting	0	25.8	34.1	32.2	28.3	total(120.34)
Overlapping	0	7.4	7.4	7.4		total(22.2)
Overlapping (fixed Δ_1)	$\Delta_1 = 8$	$\Delta_2 = 19.53$	$\Delta_3 = 35.13$	$\Delta_4 = 34.69$	$T_n = 156.63$	
Waiting	0	22.9	34.2	32.2	28.4	total(117.6)
Overlapping	0	7.4	7.4	7.4		total(22.2)

Now we consider the waiting cost, idle cost and overtime cost.

$$\begin{aligned} \min_{\mathbf{\Delta}} \quad & [c_w E_w + c_i E_i + c_o E_o] \\ \text{s.t.} \quad & W_i = (W_{i-1} + Z_{i-1} - \Delta_{i-1})^+ \\ & W_{i,i+1} = (W_{i-1} + Z_{i-1} - \Delta_{i-1} - \Delta_i)^+ \\ & E(W_{i,i+1}) \le w, i = 2, \ldots, n-1 \end{aligned} \tag{5}$$

$E_w = \sum_{i=2}^{n} E(W_i)$

$E_i = E(\sum_{i=1}^{n-1}(\Delta_i - Z_i) + W_n)$

$E_o = E(\sum_{i=1}^{n-1} \Delta_i + W_n + Z_n - T)^+$

We investigate several key questions:

When does the overlapping time constraint become active? (i.e., under what conditions does the overlapping condition take effect?) How does this constraint affect appointment scheduling? Is there an optimal approach under this constraint?

For a fixed sum $c_i + c_w$, the optimal schedule does not vary with changes in c_i or c_w.

When $c_w = 0$, problem (5) is equivalent to problem (4).

For each i, Δ_i^* increases when the threshold of waiting/overlapping time decreases.

For the optimal schedule, the expected waiting/overlapping time ($E(W_i)$, $E(W_{i,i+1})$) is increasing in i.

For the first question, we need to find the largest overlapping time.

The Wells-Riley model quantifies infection risk based on quanta emission, exposure time, masking, and ventilation. After an infected person leaves, aerosol risk drops significantly within minutes if ventilation is adequate.

Table 2: Schedule Intervals and Times

c_w	c_i	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Largest overlap time/Threshold
5	1	22	31	30	28	14	6.3462
1	1	19	29	28	28	21	5.5232
1	5	6	15	17	17	13	21.5018
1	5	6	15	18	18.61	17	18
1	5	6	15	18	21	20.23	15
1	5	6	17	19	23	24.10	12
1	5	6	18	20.83	26.03	27.97	9
1	5	7	20.25	27	28.64	30.15	6
1	5	T = 125					Infeasible < 4