UTILITY-BASED APPOINTMENT SCHEDULING

Michel Mandjes

Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands &
CWI, Amsterdam, the Netherlands &
Eurandom, Eindhoven, the Netherlands

2nd Israeli-Dutch workshop on Queueing Theory, Eurandom, September 29 - October 1 2010

Joint work with Benjamin Kemper (UvA) and Chris Klaassen (UvA, Eurandom)

SCHEDULING

When setting up an appointment schedule, it's all about balance between interests of service provider and customers:

- ⋆ if the system is frequently idle, then it is not functioning in a cost-effective manner,
- ⋆ whereas if it is virtually always busy, the customers waiting time may become substantial.

SCHEDULING

When setting up an appointment schedule, it's all about balance between interests of service provider and customers:

- $\star$ if the system is frequently idle, then it is not functioning in a cost-effective manner,
- $\star$ whereas if it is virtually always busy, the customers' waiting time may become substantial.

$\Rightarrow$ goal is to come up with a *schedule*, that is a sequence of arrival epochs.

Second question: *order* of the customers.

SETUP

$\star$ First, naïve, idea: arrival times equal sum of expected service times of *previous* customers.

This system roughly behaves as a critically loaded queue $\Rightarrow$ waiting times explode.

Good for service provider, bad for customers.

SETUP

* First, naïve, idea: arrival times equal sum of expected service times of *previous* customers.

  This system roughly behaves as a critically loaded queue $\Rightarrow$ waiting times explode.

  Good for service provider, bad for customers.

* Therefore: look for schemes that better align utilities of service provider and customers.

  With I_i idle time before i-th customer, and W_i the waiting time of i-th customer, set up schedule that sequentially minimizes utility functions $\mathbb{E}g(I_i) + \mathbb{E}h(W_i)$, for all customers i and given functions $h(\cdot)$ and $g(\cdot)$.

SETUP

- ⋆ First, naïve, idea: arrival times equal sum of expected service times of *previous* customers.

  This system roughly behaves as a critically loaded queue $\Rightarrow$ waiting times explode.

  Good for service provider, bad for customers.

- ⋆ Therefore: look for schemes that better align utilities of service provider and customers.

  With I_i idle time before i-th customer, and W_i the waiting time of i-th customer, set up schedule that sequentially minimizes utility functions $\mathbb{E}g(I_i) + \mathbb{E}h(W_i)$, for all customers i and given functions $h(\cdot)$ and $g(\cdot)$.

- ⋆ Ordering of the customers.

  For instance: in case of exponentially distributed service times, customers should be ordered such that their means (and hence also variances) increase.

SETUP

- ⋆ First, naïve, idea: arrival times equal sum of expected service times of *previous* customers.

  This system roughly behaves as a critically loaded queue $\Rightarrow$ waiting times explode.

  Good for service provider, bad for customers.

- ⋆ Therefore: look for schemes that better align utilities of service provider and customers.

  With I_i idle time before i-th customer, and W_i the waiting time of i-th customer, set up schedule that sequentially minimizes utility functions $\mathbb{E}g(I_i) + \mathbb{E}h(W_i)$, for all customers i and given functions $h(\cdot)$ and $g(\cdot)$.

- ⋆ Ordering of the customers.

  For instance: in case of exponentially distributed service times, customers should be ordered such that their means (and hence also variances) increase.

- ⋆ Examples: specific $h(\cdot)$ and $g(\cdot)$.

NAÏVE SCHEDULE

Consider sequence of jobs $B_1, \ldots, B_n$, each of random duration, assumed mutually independent. Let job i be i-th job to be scheduled.

Define standard scheduling scheme $\mathscr{S}$: arrival epoch of job i, say t_i, equals sum of expected durations of the previous jobs:

$$t_1 := 0, \quad \text{and} \quad t_i := \sum_{j=1}^{i-1} \mathbb{E}B_j,\ i = 2, \ldots, n.$$

NAÏVE SCHEDULE, ctd.

Advantage: simple!

Drawback: system essentially behaves as queue with load 1, leading to long waiting times.

Hence: for the service provider this scheme might be attractive, but for the customers it is not.

NAÏVE SCHEDULE, ctd.

Support for this claim:

Assume B_i are i.i.d. (as a random variable B) $\Longrightarrow \mathscr{S}$ can be seen as a D/G/1 queue (starting empty) with (deterministic) interarrival times equal to $b := \mathbb{E}B$. Assume $\sigma^2 := \mathbb{V}\mathrm{ar}B < \infty$.

Let W_n be waiting time of n-th customer.

Then, as $n \to \infty$,

$$\frac{\mathbb{E}W_n}{\sqrt{n}} \to \sigma\sqrt{\frac{2}{\pi}}.$$

(Remains true in the GI/G/1 setting, with $\sigma^2 := \mathbb{V}\mathrm{ar}\ A + \mathbb{V}\mathrm{ar}\ B$, where A is distributed as an interarrival time.)

NAÏVE SCHEDULE, ctd.

Main conclusion: mean waiting time under $\mathscr{S}$ grows substantially as number of customers increases.

Makespan is roughly $n\,\mathbb{E}B$, which is the best possible value (in fact, it will approximately behave as $n\,\mathbb{E}B + \sigma\sqrt{2n/\pi}$), but the waiting times increase proportionally to $\sqrt{n}$.

ADAPTED SCHEDULE, ctd.

Class of 'adapted schemes' $\mathscr{S}_\Delta$, for some $\Delta \geq 0$:

$$t_1 := 0, \quad \text{and} \quad t_i := \Delta \cdot \sum_{j=1}^{i-1} \mathbb{E}B_j,\ i = 2, \ldots, n.$$

ADAPTED SCHEDULE, ctd.

Class of 'adapted schemes' $\mathscr{S}_\Delta$, for some $\Delta \geq 0$:

$$t_1 := 0, \quad \text{and} \quad t_i := \Delta \cdot \sum_{j=1}^{i-1} \mathbb{E}B_j,\ i = 2, \ldots, n.$$

Observe

- $\star$ $\mathscr{S}_1 = \mathscr{S}$, and hence all previous results relate to the case $\Delta = 1$.
- $\star$ Makespan is reduced (compared to $\mathscr{S}$) when picking $\Delta \in [0, 1)$;

  in extreme case of $\Delta = 0$, all customers arrive at time 0, thus minimizing the expected makespan (at the expense of the waiting time of the customers).
- $\star$ Mean delays are reduced (relative to $\mathscr{S}$) when picking $\Delta > 1$ (at the expense of idle time of the server);

  corresponding D/G/1 queue is stable, i.e., it has a proper steady-state distribution.

ADAPTED SCHEDULE, ctd.

Drawback: scheme only depends on *mean* service times.

ADAPTED SCHEDULE, ctd.

Drawback: scheme only depends on *mean* service times.

Depending on the shape of the service time distributions, the mean waiting time may wildly vary.

Put differently: for given Δ performance of schedule critically depends on service time distribution.

ADAPTED SCHEDULE, ctd.

Drawback: scheme only depends on *mean* service times.

Depending on the shape of the service time distributions, the mean waiting time may wildly vary.

Put differently: for given Δ performance of schedule critically depends on service time distribution.

Therefore: need for schedules that better balance interests of customers and provider.

RISK FUNCTIONS

Key notion: 'risk', measures aggregate disutility of the server and client.

More specifically: risk associated with i-th arrival depends on the distribution of waiting time W_i of the i-th client, and idle time I_i prior to the arrival of this i-th client.

RISK FUNCTIONS

Key notion: 'risk', measures aggregate disutility of the server and client.

More specifically: risk associated with i-th arrival depends on the distribution of waiting time W_i of the i-th client, and idle time I_i prior to the arrival of this i-th client.

Choose nondecreasing functions $g(\cdot)$ and $h(\cdot)$ with $g(0) = h(0) = 0$, and define risk at i-th arrival as

$$R_i^{(g,h)}(t_1, \ldots, t_i) = \mathbb{E}g(I_i) + \mathbb{E}h(W_i).$$

$g(\cdot)$ and $h(\cdot)$ determine weight given to idle and waiting time respectively; risk depends on the schedule up to the i-th appointment time.

RISK FUNCTIONS, ctd.

Risk function:

$$R_i^{(g,h)}(t_1, \ldots, t_i) = \mathbb{E}g(I_i) + \mathbb{E}h(W_i).$$

I_i and W_i cannot be both positive; natural to introduce loss function

$$\ell(x) = g(-x)\mathbf{1}_{[x<0]} + h(x)\mathbf{1}_{[x>0]}, \quad x \in \mathbb{R},$$

nonincreasing on $(-\infty, 0]$ and nondecreasing on $[0, \infty)$ with $\ell(0) = 0$.

RISK FUNCTIONS, ctd.

Risk function:

$$R_i^{(g,h)}(t_1, \ldots, t_i) = \mathbb{E}g(I_i) + \mathbb{E}h(W_i).$$

I_i and W_i cannot be both positive; natural to introduce loss function

$$\ell(x) = g(-x)\mathbf{1}_{[x<0]} + h(x)\mathbf{1}_{[x>0]}, \quad x \in \mathbb{R},$$

nonincreasing on $(-\infty, 0]$ and nondecreasing on $[0, \infty)$ with $\ell(0) = 0$.

Hence

$$R_i^{(g,h)}(t_1, \ldots, t_i) = \mathbb{E}g(I_i) + \mathbb{E}h(W_i) = \mathbb{E}\ell(W_i - I_i),$$

and we define the risk at the i-th arrival with loss function $\ell(\cdot)$ as

$$R_i^{(\ell)}(t_1, \ldots, t_i) = \mathbb{E}\ell(W_i - I_i).$$

RISK FUNCTIONS, ctd.

Goal:

sequentially optimize appointment times,

i.e., optimize the choice of t_i, given the appointment times $0 = t_1, \ldots, t_{i-1}$.

RISK FUNCTIONS, ctd.

Observe: both I_1 and W_1 vanish.

Due to Lindley recursion

$$I_i = \max\{t_i - t_{i-1} - W_{i-1} - B_{i-1}, 0\}$$

and

$$W_i = \max\{W_{i-1} + B_{i-1} - t_i + t_{i-1}, 0\}.$$

Hence

$$W_i - I_i = W_{i-1} + B_{i-1} - t_i + t_{i-1}.$$

RISK FUNCTIONS, ctd.

Let $S_i := W_i + B_i$ denote sojourn time of the i-th customer, with density $f_{S_i}(\cdot)$ and distribution function $F_{S_i}(\cdot)$.

In addition, let $x_{i-1} := t_i - t_{i-1}$ be the time between the $(i-1)$-st and i-th arrival.

Then we may write

$$W_i - I_i = S_{i-1} - x_{i-1}$$

and

$$R_i^{(\ell)}(t_1, \ldots, t_{i-1}, t_{i-1} + x_{i-1}) = \mathbb{E}\ell(S_{i-1} - x_{i-1}).$$

RISK FUNCTIONS, ctd.

General condition for the sequential optimization of the risk at the i-th arrival.

Theorem. Let $\ell(\cdot)$ be a nonnegative convex function on $\mathbb{R}$ with $\ell(0) = 0$.

Then $\ell(\cdot)$ is a loss function, i.e., it is nonincreasing on $(-\infty, 0]$ and nondecreasing on $[0, \infty)$ with $\ell(0) = 0$, and it is absolutely continuous with derivative $\ell'(\cdot)$.

Let S be a random variable with a density with respect to Lebesgue measure and let $\mathbb{E}\ell(S - x)$ and $\mathbb{E}\ell'(S - x)$ be finite for all $x \in \mathbb{R}$.

Then

$$\inf_{x \in \mathbb{R}} \mathbb{E}\ell(S - x)$$

is attained at $x^\star$ if and only if

$$\mathbb{E}\ell'(S - x^\star) = 0$$

holds.

RISK FUNCTIONS, ctd.

Proof Risk function $R := \mathbb{E}g(I) + \mathbb{E}h(W)$ can be evaluated as

$$\int_0^\infty g(s) f_I(s) \mathrm{d}s + \int_0^\infty h(s) f_W(s) \mathrm{d}s$$

for any client i; here $f_I(\cdot)$ and $f_W(\cdot)$ are the densities of I and W.

Recalling $S_{i-1} = W_{i-1} + B_{i-1}$ and $x_{i-1} = t_i - t_{i-1}$, rewrite R_i as

$$\Phi(x_{i-1}) := \int_0^{x_{i-1}} g(x_{i-1} - s) f_{S_{i-1}}(s) \mathrm{d}s + \int_{x_{i-1}}^\infty h(s - x_{i-1}) f_{S_{i-1}}(s) \mathrm{d}s.$$

Limits of integration and integrands are functions of the interarrival time x_{i-1} — apply Leibniz's rule:

$$\Phi'(x) = g(0) f_{S_{i-1}}(x) - h(0) f_{S_{i-1}}(x) + \int_0^x g'(x-s) f_{S_{i-1}}(s) \mathrm{d}s - \int_x^\infty h'(s-x) f_{S_{i-1}}(s) \mathrm{d}s,$$

yields the stated. $\square$

EXAMPLE: LINEAR

Consider *linear risks*:

$$
\begin{aligned}
R_i^{(a)}(t_1, \ldots, t_{i-1}, t_{i-1} + x) &:= \mathbb{E} I_i + \mathbb{E} W_i \\
&= \mathbb{E} \left| S_{i-1} - x \right|.
\end{aligned}
$$

EXAMPLE: LINEAR

Consider *linear risks*:

$$
\begin{aligned}
R_i^{(a)}(t_1, \ldots, t_{i-1}, t_{i-1} + x) &:= \mathbb{E} I_i + \mathbb{E} W_i \\
&= \mathbb{E}\,|S_{i-1} - x|.
\end{aligned}
$$

According to our theorem this expression is minimized for any $x > 0$ satisfying

$$
\int_0^x f_{S_{i-1}}(s)\mathrm{d}s = \int_x^\infty f_{S_{i-1}}(s)\mathrm{d}s.
$$

This implies that $x_{i-1}^\star$ should equal a *median* of S_{i-1}: $x_{i-1}^\star = F_{S_{i-1}}^{-1}(\frac{1}{2})$.

(Reminiscence with *newsvendor problem*)

EXAMPLE: LINEAR

Consider *linear risks*:

$$\begin{aligned} R_i^{(a)}(t_1, \ldots, t_{i-1}, t_{i-1} + x) &:= \mathbb{E} I_i + \mathbb{E} W_i \\ &= \mathbb{E}\,|S_{i-1} - x|. \end{aligned}$$

According to our theorem this expression is minimized for any $x > 0$ satisfying

$$\int_0^x f_{S_{i-1}}(s)\mathrm{d}s = \int_x^\infty f_{S_{i-1}}(s)\mathrm{d}s.$$

This implies that $x_{i-1}^\star$ should equal a *median* of S_{i-1}: $x_{i-1}^\star = F_{S_{i-1}}^{-1}(\frac{1}{2})$.

(Reminiscence with *newsvendor problem*)

Hence, optimal to choose t_i (given t_1 up to t_{i-1}) according to the schedule $\mathscr{T}$ given by

$$t_i := t_{i-1} + F_{S_{i-1}}^{-1}\left(\frac{1}{2}\right).$$

EXAMPLE: LINEAR, ctd.

Similar loss functions can be treated in the same way.

Example: $R_i^{(m)}(t_1, \ldots, t_i) := \mathbb{E} \max\{I_i, W_i\}$.

The identity

$$\max\{0, x - S\} + \max\{0, S - x\} = |S - x| = \max\{\max\{0, x - S\}, \max\{0, S - x\}\}$$

immediately implies that $\mathscr{T}$ also sequentially minimizes the risk $R_i^{(m)}(t_1, \ldots, t_i)$, for $i = 1, \ldots, n$.

EXAMPLE: QUADRATIC

Now consider quadratic risks:

$$R_i^{(q)}(t_1, \ldots, t_i) := \mathbb{E}I_i^2 + \mathbb{E}W_i^2, \quad i = 2, \ldots, n.$$

Define schedule $\mathscr{V}$ through

$$t_1 := 0, \quad \text{and} \quad t_i := \sum_{j=1}^{i-1} \mathbb{E}S_j,\ i = 1, \ldots, n.$$

We can verify that $\mathscr{V}$ is optimal by applying our theorem; we however add an alternative, insightful approach.

EXAMPLE: QUADRATIC, ctd.

Observe that $W_1 = 0$ and $I_1 = 0$

Also,

$$I_i^2 + W_i^2 = (t_i - t_{i-1} - W_{i-1} - B_{i-1})^2 = (t_i - t_{i-1} - S_{i-1})^2.$$

Now minimize, for given t_{i-1}, risk of customer i:

$$\min_{t_i} R_i^{(q)}(t_1, \ldots, t_i) = \min_{t_i} \mathbb{E}(t_i - t_{i-1} - S_{i-1})^2 = \mathbb{V}\mathrm{ar}\, S_{i-1},$$

with $t_i - t_{i-1} = \mathbb{E} S_{i-1}$.

The schedule $\mathscr{V}$ sequentially minimizes the risk $R_i^{(q)}(t_1, \ldots, t_i)$, for $i = 1, \ldots, n$.

ORDERING

Main contribution here:

consider n customers with independent service times $B_1, \ldots, B_n$, and let B_i be distributed as $\sigma_i B_1$ for $i = 1, \ldots, n$, assuming $\sigma_1 = 1 \leq \sigma_2 \leq \ldots \leq \sigma_n$.

ORDERING

Main contribution here:

consider n customers with independent service times $B_1, \ldots, B_n$, and let B_i be distributed as $\sigma_i B_1$ for $i = 1, \ldots, n$, assuming $\sigma_1 = 1 \leq \sigma_2 \leq \ldots \leq \sigma_n$.

Define an *ordering* $N(\cdot)$ as a mapping that bijectively projects $\{1, \ldots, n\}$ onto $\{1, \ldots, n\}$.

ORDERING

Main contribution here:

consider n customers with independent service times $B_1, \ldots, B_n$, and let B_i be distributed as $\sigma_i B_1$ for $i = 1, \ldots, n$, assuming $\sigma_1 = 1 \leq \sigma_2 \leq \ldots \leq \sigma_n$.

Define an *ordering* $N(\cdot)$ as a mapping that bijectively projects $\{1, \ldots, n\}$ onto $\{1, \ldots, n\}$.

Then, in order to find the order that sequentially optimizes the risks, the mapping $N(\cdot)$ should be such that the σ_i are nondecreasing, given that for any order the schedule is in accordance with our theorem.

ORDERING, ctd.

Proof Write $W_i - I_i = W_{i-1} + B_{i-1} - (t_i - t_{i-1})$.

Applying our theorem: $R_i = \inf_{x_{i-1}} \mathbb{E}\ell(W_{i-1} + B_{i-1} - x_{i-1})$.

For any optimal interarrival time, we study the risk function $\psi(\cdot)$ in terms of scale parameter σ of service time distribution:

$$\psi(\sigma) = \inf_x \mathbb{E}\ell(W + \sigma B - x) = \mathbb{E}\ell(W + \sigma B - x^\star_\sigma),$$

with $B \equiv B_1$, and $x^\star_\sigma$ the optimizing x as a function of σ.

Notice that we have proved our claim if we can show that $\psi(\sigma)$ increases in σ.

ORDERING, ctd.

First order condition states that

$$\mathbb{E}\left(\frac{\partial}{\partial x}\ell(W+\sigma B-x)\right)\Big|_{x=x^\star_\sigma} = -\mathbb{E}\ell'(W+\sigma B-x^\star_\sigma) = 0;$$

W and B are independent.

ORDERING, ctd.

First order condition states that

$$\mathbb{E}\left(\frac{\partial}{\partial x}\ell(W+\sigma B-x)\right)\Bigg|_{x=x^\star_\sigma} = -\mathbb{E}\ell'(W+\sigma B-x^\star_\sigma)=0;$$

W and B are independent.

Lots of computations:

$$\begin{aligned}
\psi'(\sigma) &= \mathbb{E}[\ell'(W+\sigma B-x^\star_\sigma)(B-\dot{x}^\star_\sigma)] \\
&= \mathbb{E}[\ell'(W+\sigma B-x^\star_\sigma)B] = \mathbb{E}[\mathbb{E}[\ell'(W+\sigma B-x^\star_\sigma)B \mid W]] \\
&= \mathbb{E}\left[\mathbb{E}\left[\{\ell'(W+\sigma B-x^\star_\sigma) - \mathbb{E}\ell'(W+\sigma B-x^\star_\sigma) \mid W\}B \mid W\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\ell'(W+\sigma B-x^\star_\sigma)B - \mathbb{E}[\ell'(W+\sigma B-x^\star_\sigma)|W]B \mid W\right]\right] \\
&= \mathbb{E}\left[\mathbb{C}\mathrm{ov}\left[\ell'(W+\sigma B-x^\star_\sigma), B \mid W\right]\right].
\end{aligned}$$

Since $\ell(\cdot)$ is strictly convex, $\ell'(W+\sigma B-x^\star_\sigma)$ is increasing in B, and therefore

$$\mathbb{C}\mathrm{ov}(\ell'(w+\sigma B-x^\star_\sigma), B) > 0$$

for any $w \geq 0$. $\square$

ADDITIONAL FEATURES

We can also deal with

- $\star$ Additional urgent arrivals;
- $\star$ No shows.

STEADY-STATE

Effect of scheduling policies $\mathscr{T}$ and $\mathscr{V}$ by considering the situation of i.i.d. jobs, and the number of jobs n being large.

Goal: limiting interarrival time for both scheduling policies.

Assume the jobs are exponential with mean $1/\mu$, so that the queue under consideration is an D/M/1. Let x be the interarrival time between two subsequent jobs.

Then distribution of the steady-state waiting time W is given through

$$\mathbb{P}(W > y) = \sigma_x e^{-\mu(1-\sigma_x)y}, \quad y > 0,$$

where $\sigma \equiv \sigma_x$ is the unique solution in $(0,1)$ of $e^{-\mu(1-\sigma)x} = \sigma$.

STEADY-STATE

First consider linear loss function and strategy $\mathscr{T}$.

Then it turns out that

$$G(y) := \mathbb{P}(W + B \leq y) = 1 - e^{-\mu(1-\sigma_x)y}, \quad y > 0.$$

It follows directly that

$$G^{-1}\left(\frac{1}{2}\right) = \frac{\log 2}{\mu(1-\sigma_x)}.$$

We find for the optimal interarrival time $x^\star$

$$\sigma_{x^\star} = \frac{1}{2}, \quad \text{and} \quad x^\star = \frac{1}{\mu} \cdot 2\log 2.$$

STEADY-STATE

Now focus on quadratic loss function and policy $\mathscr{V}$. Then

$$\mathbb{E}W + \mathbb{E}B = \frac{\sigma_x}{\mu(1-\sigma_x)} + \frac{1}{\mu} = \frac{1}{\mu(1-\sigma_x)}.$$

Straightforward calculations, with $x^\star$ being the optimal interarrival time,

$$\sigma_{x^\star} = \frac{1}{e}, \quad \text{and} \quad x^\star = \frac{1}{\mu} \cdot \frac{e}{e-1}.$$

STEADY-STATE

Now focus on quadratic loss function and policy $\mathscr{V}$. Then

$$\mathbb{E}W + \mathbb{E}B = \frac{\sigma_x}{\mu(1-\sigma_x)} + \frac{1}{\mu} = \frac{1}{\mu(1-\sigma_x)}.$$

Straightforward calculations, with $x^\star$ being the optimal interarrival time,

$$\sigma_{x^\star} = \frac{1}{e}, \quad \text{and} \quad x^\star = \frac{1}{\mu} \cdot \frac{e}{e-1}.$$

As $e/(e-1) \approx 1.5820$ and $2\log 2 \approx 1.3863$: under the quadratic loss function the scheduling is somewhat more 'defensive' than under the linear loss function.

CONCLUDING REMARKS

- $\star$ Method proposed balances the customers' and the provider's interests;
- $\star$ Non-limiting regime (i.e., n not large) raises computational questions — but usually rapid convergence to steady-state;
- $\star$ Ordering problem solved if jobs are from same scale-family.