Elective Surgery Sequencing and Scheduling under Uncertainty

Jin QI[1]
Joint work with Xiaojin Fu[2], Chen Yang[3] and Han Ye[4]

[1]Hong Kong University of Science and Technology, [2]Huawei Technologies, HK
[3]Zhejiang University, [4]Lehigh University

Why OR Scheduling Matters

Operating rooms are a high-stakes bottleneck.

- ORs use **scarce resources**: rooms, equipment, surgeons, anesthesiologists, and nursing teams.
- ORs account for more than 40% of **hospital cost and revenue**. (Denton et al. 2007, Freeman et al. 2016)
- More than 60% of hospital admissions require **surgical intervention**. (Guerriero and Guido 2011, Fugener 2017)

Scheduling quality affects stakeholders differently.

- Patients care about waiting and uncertainty.
- Surgical teams care about overtime and disruption.
- Managers care about utilization and quality.

Core tension

Tight schedules improve utilization, but amplify delay risk under uncertain durations.

Why OR Scheduling Matters

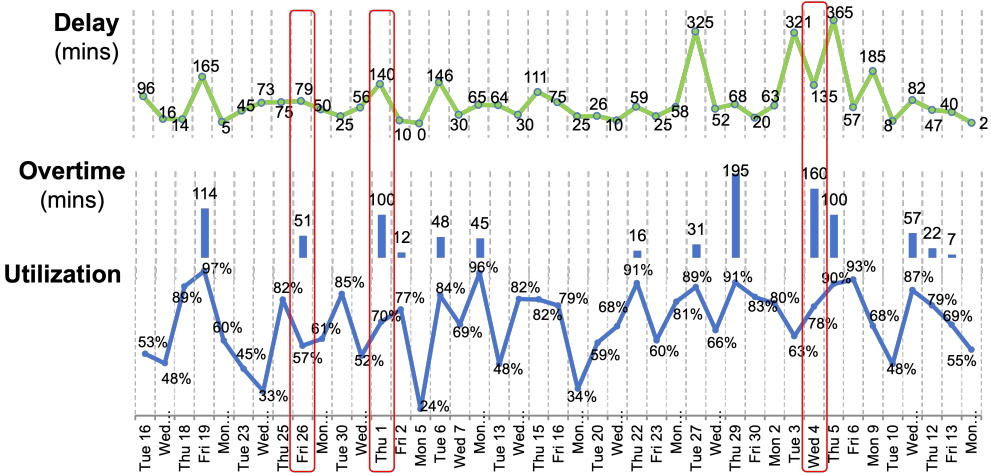
Significance of scheduling:

- Operating rooms (ORs) are valuable and limited.
- Tremendous uncertainty in surgical duration.
- High rates of delay and overtime in practice.

Surgery types:

- **Elective surgery:** can be scheduled several weeks or more in advance.
- **Emergency surgery:** requires immediate surgical intervention, and is usually conducted in specialized ORs.

Empirical Evidence



Working hour: 8:30-16:30

Challenge 1: What is the Objective?

Empirical evidence from prior studies

- In the US, only **27%** of surgeries start on schedule in one cited perioperative workflow study. (Denton et al. 2007, Doebbeling et al. 2012)
- At a German university hospital, **87%** of elective cases started more than 10 minutes away from the planned start time; **26%** started **early**. (Balzer et al. 2017)
- In our real data, the average OR delay is 41.5 mins; 45% of surgeries have delays $\geq$ 30 mins, 24% have delays $\geq$ 60 mins.

Implication for modeling

How should we balance the delay and idle time? Is the average enough?

Challenge 2: How to Describe Uncertainty

Suppose three surgeries have the same mean but different uncertainty

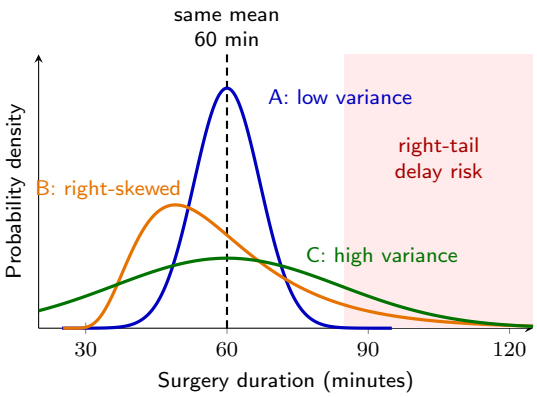
Surgery	Mean duration	Variance	Tail behavior
A	60 min	Low	nearly symmetric
B	60 min	Medium	strongly right-skewed
C	60 min	High	symmetric

Put B first?

Variance is lower, but the long right tail of B can create **cascading delay** for all following surgeries.

Put C first?

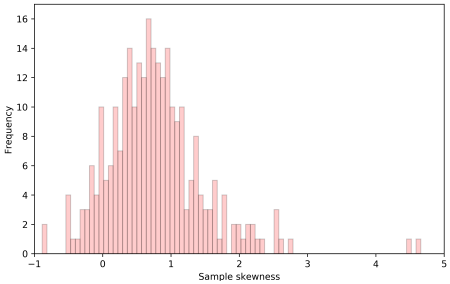
Variance is higher, but a more symmetric distribution may be **easier to protect** with planned buffer.



Challenge 2: How to Describe Uncertainty?

Duration uncertainty has shape.

- Surgeries with the same variance can have very different right-tail behavior.
- In our real data, **88.7%** of surgery types with more than 10 samples are **right-skewed**.
- Right-skewness matters because long procedures propagate delay to every following surgery.

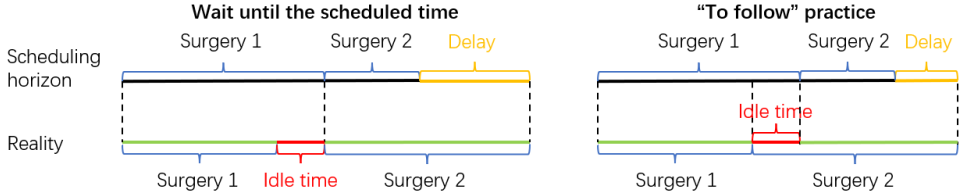


Sequencing question

Should we sort surgeries only by variance, or should the direction of uncertainty matter?

Challenge 3: "To Follow" Practice

"To follow": a surgery is conducted immediately after the preceding one.



- Widely observed. (Pinedo 2012, van Essen et al. 2012, Balzer et al. 2017, Bai et al. 2020).
- Real data: if the preceding surgery ends early, 65% of surgeries start early.

The preceding surgery ends early.
Negative values: earlier than the scheduled time.

Range of preceding delay (min)		> 20	$(10, 20]$	$(0, 10]$	$(-10, 0]$	$(-20, -10]$	$(-30, -20]$	$(-40, -30]$	< -40
Start time delay (min)	Q1	49.3	20.2	10.2	0.7	-8.7	-18.3	-24.9	-40.5
	Median	70.2	25.0	15.6	5.6	-0.3	-10.0	-18.4	-30.0
	Q3	104.3	30.7	23.6	13.9	5.8	-1.8	-4.0	-14.3
	Mean	81.9	28.5	20.0	10.9	2.7	-4.2	-12.5	-25.6
Proportion (%)		60.3	9.5	7.6	6.9	5.4	3.7	2.3	4.3

Challenge 3: "To Follow" Practice

"To follow" is not just a modeling convenience.

- Surgical teams often **operate sequentially in a block**, creating a natural incentive to continue immediately.
- Patients are usually asked to **arrive before** the scheduled surgery time. (Guda et al. 2016, Bai et al. 2020)
- The same practice has been observed across multiple hospital settings. (van Essen et al. 2012, Balzer et al. 2017, Bai et al. 2020)

Conventional assumption

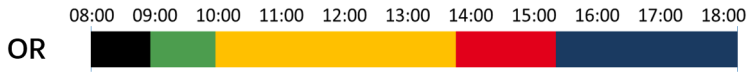
A surgery cannot start before its scheduled time.

Observed practice

If the preceding surgery finishes early and the next patient is ready, the next surgery can start early.

Research Question

Elective surgery sequencing and scheduling in a single OR



- **Input**: surgeries and their historical data.
- **Goal**: optimize the risk of the **delay** and **idle time**.
- **Settings and assumptions**:
  - "To follow" practice.
  - Uncertain and independent surgical durations.
- **Decisions:**
  - Sequencing decision (Binary)
  - Scheduling decision (Continuous)

Literature Review

Methodology

- Stochastic programming: Denton and Gupta (2003), Denton et al. (2007), Freeman et al. (2016), Guda et al. (2016), Sauré et al. (2020), Yiu et al. (2024).
- Distributionally robust optimization (DRO): Kong et al. (2013), Mak et al. (2014), Jiang et al. (2019), Shehadeh et al. (2022).

Performance measures

- Expected delay, idle time and overtime: Denton et al. (2007), Kong et al. (2013), Mak et al. (2014), Jiang et al. (2019).
- Chance constraints on acceptable overtime levels: Shylo et al. (2012).
- Overtime risk based on exponential disutility: Qi (2016), Zhang et al. (2020).

Sequencing decision

- Smallest variance first (SVF): Mak et al. (2014), Baker (2014), Otten et al. (2019).
- Other heuristics: Mancilla and Storer (2012), Zacharias and Pinedo (2014).
- Optimality condition for SVF: dilation order of durations (Guda et al. 2016).

Gaps and Contributions

Gaps

- "To follow" practice.
- Limited risk evaluation of delay and idle time.
- The joint problem of sequencing and scheduling is challenging.

Contributions

- **Punctuality index**: a new performance measure to evaluate the risk of delay and idle time.
- A novel and tractable model for both **sequencing and scheduling**.
- Two simple **heuristics**: variance and directional deviations.
- **Case study** using data from a Singapore hospital.

Notation

Decision variables and feasible regions

- x_{in}: Binary decision of whether surgery n is assigned at position i.
- s_i: the scheduled start time of the surgery at position i.
- $\mathcal{S}, \mathcal{X}$: feasible sets of decision variables $\boldsymbol{s}$ and $\boldsymbol{x}$.

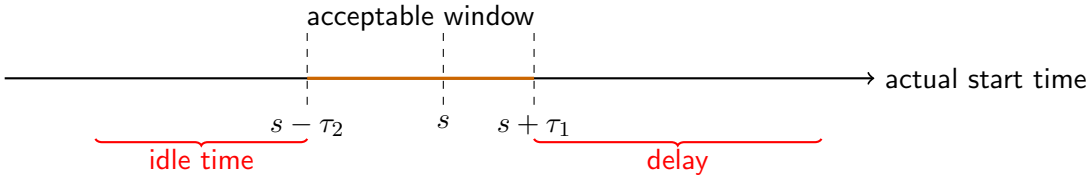
$$\mathcal{S} = \left\{ \boldsymbol{s} \in \Re_+^N \, | s_j \leq s_{j+1}, \; j \in [N-1] \right\},$$

$$\mathcal{X} = \left\{ \boldsymbol{X} \in \{0,1\}^{N \times N} \middle| \sum_{n \in [N]} x_{jn} = 1, j \in [N], \sum_{j \in [N]} x_{jn} = 1, n \in [N] \right\}.$$

Parameters

- $\tilde{t}_n$: the uncertain duration of surgery $n \in [N]$.
- τ_1, τ_2: tolerance levels for delay and idle time. For surgery at position j, we expect its actual start time to lie within $[s_j - \tau_2, s_j + \tau_1]$.

Risk-neutral Model



Risk-neutral Model: minimizes the weighted sum of expected delay and idle time.

$$\sum_{j\in[N-1]} \left(\omega_1 \underbrace{\mathbb{E}\left[\left(\sum_{i\in[j]} \sum_{n\in[N]} x_{in}\tilde{t}_n - s_{j+1} - \tau_1 \right)^+ \right]}_{\text{Expected delay of surgery at position } j+1} + \omega_2 \underbrace{\mathbb{E}\left[\left(s_{j+1} - \sum_{i\in[j]} \sum_{n\in[N]} x_{in}\tilde{t}_n - \tau_2 \right)^+ \right]}_{\text{Expected idle time of surgery at position } j+1} \right).$$

actual start time

actual start time

What We Need From a Risk Measure

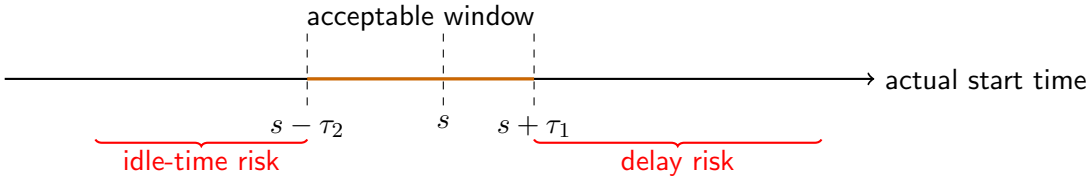
The risk-neutral model helps, but it is not enough for our goal.

- It measures **average** delay and **average** idle time.
- It treats many small violations and large violations **equally**.
- It is **harder to compute** because decisions appear inside $(\cdot)^+$.

Desired properties

- Reflect both the chance and the size of violations.
- Penalize extreme outcomes (tail risk) for risk-averse decisions.
- Keep enough structure so the combined sequencing-scheduling problem stays tractable.

Punctuality Index: Intuition



- The schedule is acceptable if the actual start time lies in $[s - \tau_2, s + \tau_1]$.
- Starting too early means resources are ready but the planned schedule has too much slack.
- Starting too late means patients and downstream surgeries wait.

Role of the Punctuality Index

Converts this two-sided risk into an optimization-friendly quantity.

Worst-case Certainty Equivalent

For an uncertain variable $\tilde{t}$ with ambiguity set $\mathbb{F}$, the worst-case certainty equivalent is

$$C_\alpha(\tilde{t}) = \begin{cases} \alpha \ln \sup_{\mathbb{P}\in\mathbb{F}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{t}}{\alpha}\right)\right], & \text{if } \alpha > 0 \\ \lim_{\gamma\downarrow 0} C_\gamma(\tilde{t}), & \text{if } \alpha = 0 \end{cases}$$

worst-case distribution in $\mathbb{F}$

Remark: a concept from exponential disutility theory.

$$\exp\left(\frac{C_\alpha(\tilde{t})}{\alpha}\right) = \sup_{\mathbb{P}\in\mathbb{F}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{t}}{\alpha}\right)\right] \Longrightarrow C_\alpha(\tilde{t}) = \alpha \ln \sup_{\mathbb{P}\in\mathbb{F}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{t}}{\alpha}\right)\right]$$

Properties

- Monotonicity: $C_\alpha(\tilde{t})$ is decreasing in α;
- Convexity: $C_\alpha(\tilde{t})$ is jointly convex in α and $\tilde{t}$;
- Additivity: $C_\alpha(\tilde{t}_1 + \tilde{t}_2) = C_\alpha(\tilde{t}_1) + C_\alpha(\tilde{t}_2)$ if $\tilde{t}_1$ is independent with $\tilde{t}_2$.

Punctuality Index

Punctuality Index $\rho_{s-\tau_2,s+\tau_1}(\tilde{t}) : \mathcal{V} \to [0, +\infty]$

$$\rho_{s-\tau_2,s+\tau_1}(\tilde{t}) = \inf \left\{ \omega_1\alpha_1 + \omega_2\alpha_2 \,\middle|\, \underbrace{C_{\alpha_1}(\tilde{t}) \leq s + \tau_1}_{\text{risk of delay}}, \underbrace{C_{\alpha_2}(-\tilde{t}) \leq -(s - \tau_2)}_{\text{risk of idle time}}, \alpha_1, \alpha_2 \geq 0 \right\},$$

or ∞ if no solution exists.

Punctuality Index

$$\rho_{s-\tau_2,s+\tau_1}(\tilde{t}) = \inf\left\{\omega_1\alpha_1 + \omega_2\alpha_2 \,\middle|\, C_{\alpha_1}(\tilde{t}) \le s+\tau_1, C_{\alpha_2}(-\tilde{t}) \le -(s-\tau_2), \alpha_1,\alpha_2 \ge 0\right\},$$

Properties

- Punctuality satisficing: $\rho_{s-\tau_2,s+\tau_1}(\tilde{t}) = 0$ if and only if $\mathbb{P}(s-\tau_2 \le \tilde{t} \le s+\tau_1) = 1$ for all $\mathbb{P} \in \mathbb{F}$;
- Idle time intolerance: if $\inf_{\mathbb{P}\in\mathbb{F}} \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] < s-\tau_2$, then $\rho_{s-\tau_2,s+\tau_1}(\tilde{t}) = \infty$;
- Delay intolerance: if $\sup_{\mathbb{P}\in\mathbb{F}} \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] > s+\tau_1$, then $\rho_{s-\tau_2,s+\tau_1}(\tilde{t}) = \infty$;
- Punctuality guarantee: let (α_1^*, α_2^*) denote the optimal solution in $\rho_{s-\tau_2,s+\tau_1}(\tilde{t})$. For any violation $\theta > 0$,

$$\sup_{\mathbb{P}\in\mathbb{F}} \mathbb{P}(\tilde{t} \ge s+\tau_1+\theta) \le \exp\left(-\frac{\theta}{\alpha_1^*}\right), \quad \sup_{\mathbb{P}\in\mathbb{F}} \mathbb{P}(\tilde{t} \le s-\tau_2-\theta) \le \exp\left(-\frac{\theta}{\alpha_2^*}\right).$$

Formulation

$$\rho_{s-\tau_2,s+\tau_1}(\tilde{t}) = \inf\left\{\omega_1\alpha_1 + \omega_2\alpha_2 \,\middle|\, C_{\alpha_1}(\tilde{t}) \le s+\tau_1, C_{\alpha_2}(-\tilde{t}) \le -(s-\tau_2), \alpha_1,\alpha_2 \ge 0\right\},$$

Minimize the sum of Punctuality Indices $\sum_{j=2}^{N} \rho_{s_j-\tau_2,s_j+\tau_1}\left(\sum_{i\in[j-1]}\sum_{n\in[N]} x_{in}\tilde{t}_n\right)$

$$\inf_{\boldsymbol{X}\in\mathcal{X},\boldsymbol{s}\in\mathcal{S},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2\ge\mathbf{0}} \sum_{j\in[N-1]} (\omega_1\alpha_{1j} + \omega_2\alpha_{2j}) \longrightarrow \text{sum of risk tolerances } \alpha$$

risk of delay

$$\text{s.t.} \quad C_{\alpha_{1j}}\left(\sum_{i\in[j]}\sum_{n\in[N]} x_{in}\tilde{t}_n\right) \le s_{j+1}+\tau_1, \qquad j\in[N-1],$$

$$C_{\alpha_{2j}}\left(-\sum_{i\in[j]}\sum_{n\in[N]} x_{in}\tilde{t}_n\right) \le -(s_{j+1}-\tau_2), \quad j\in[N-1].$$

risk of idle time

Additivity:

$$C_{\alpha_{1j}}\left(\sum_{i\in[j]}\sum_{n\in[N]} x_{in}\tilde{t}_n\right) = \sum_{i\in[j]}\sum_{n\in[N]} x_{in}C_{\alpha_{1j}}\left(\tilde{t}_n\right) \le s_{j+1}+\tau_1,$$

$$C_{\alpha_{2j}}\left(-\sum_{i\in[j]}\sum_{n\in[N]} x_{in}\tilde{t}_n\right) = \sum_{i\in[j]}\sum_{n\in[N]} x_{in}C_{\alpha_{2j}}\left(-\tilde{t}_n\right) \le -(s_{j+1}-\tau_2).$$

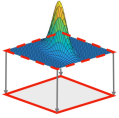
Distributionally Robust Optimization

From classical stochastic programming to DRO:

$$C_\alpha(\tilde{t}) = \begin{cases} \alpha \ln \sup_{\mathbb{P}\in\mathbb{F}} \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{t}}{\alpha}\right)\right], & \text{if } \alpha > 0 \\ \lim_{\gamma\downarrow 0} C_\gamma(\tilde{t}), & \text{if } \alpha = 0 \end{cases}$$

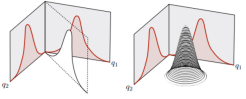
Ambiguity Set $\mathbb{F}$:

Support

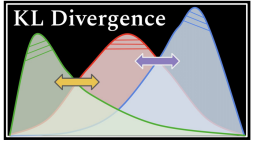


Moments

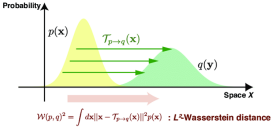
Structural Information



Statistical Divergence



Optimal Transport Cost



Special Case 1: Sequencing Decision

Dilation order: $\tilde{t}_1$ is smaller than $\tilde{t}_2$ in the dilation order, denoted as $\tilde{t}_1 \leq_{dil} \tilde{t}_2$, if $\mathbb{E}[\phi(\tilde{t}_1 - \mathbb{E}[\tilde{t}_1])] \leq \mathbb{E}[\phi(\tilde{t}_2 - \mathbb{E}[\tilde{t}_2])]$ holds for any convex function $\phi(\cdot)$.

Proposition

If there exists a sequence $(k_1, k_2, \ldots, k_N)$ such that $\tilde{t}_{k_1} \leq_{dil} \tilde{t}_{k_2} \leq_{dil} \ldots \leq_{dil} \tilde{t}_{k_N}$, then the sequence $(k_1, k_2, \ldots, k_N)$ is optimal.

- $\tilde{t}_n = \mathbb{E}_{\mathbb{P}}[\tilde{t}_n] + k_n \tilde{u}_n$, $\tilde{u}_n$ are i.i.d.
- Smallest Variance First (SVF): a special case with $\phi(x) = x^2$.

Special Case 2: Scheduling Decision

Given sequencing decisions, if $\omega_1 = 0$ (only idle time), the optimal scheduled start time is

$$s^*_{j+1} = \sum_{i\in[j]} \sum_{n\in[N]} x_{in} \sup_{\mathbb{P}_n \in \mathbb{F}_n} \mathbb{E}_{\mathbb{P}_n}\left[\tilde{t}_n\right] - \tau_1, \quad j \in [N-1].$$

If $\omega_2 = 0$ (only delay), the optimal scheduled start time is

$$s^*_{j+1} = \sum_{i\in[j]} \sum_{n\in[N]} x_{in} \inf_{\mathbb{P}_n \in \mathbb{F}_n} \mathbb{E}_{\mathbb{P}_n}\left[\tilde{t}_n\right] + \tau_2, \quad j \in [N-1].$$

Special Case 3: Gaussian Case

When the surgery durations $(\tilde{t}_n)_{n\in[N]} \sim N(\mu_n, \sigma_n^2)$

- Optimal sequencing: SVF;
- Optimal scheduling: $s^*_{j+1} = \sum_{i\in[j]} \sum_{n\in[N]} x^*_{in}\mu_n + \dfrac{\sqrt{\omega_1}\tau_2 - \sqrt{\omega_2}\tau_1}{\sqrt{\omega_1} + \sqrt{\omega_2}}, \quad j \in [N-1].$

Solution Procedure

Proposition

The whole model is a mixed-integer convex optimization problem.

The subproblem (scheduling): given $\boldsymbol{X}$, the problem is convex.

$$
\begin{aligned}
f(\boldsymbol{X}) = \inf_{\boldsymbol{s}\in\mathcal{S},\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2\geq\mathbf{0}} & \sum_{j\in[N-1]} (\omega_1\alpha_{1j} + \omega_2\alpha_{2j}) \\
\text{s.t.} \quad & \sum_{i\in[j]}\sum_{n\in[N]} x_{in} C_{\alpha_{1j}}\left(\tilde{t}_n\right) \leq s_{j+1} + \tau_1, \qquad j\in[N-1], \\
& \sum_{i\in[j]}\sum_{n\in[N]} x_{in} C_{\alpha_{2j}}\left(-\tilde{t}_n\right) \leq -(s_{j+1} - \tau_2), \quad j\in[N-1].
\end{aligned}
$$

The master problem (sequencing): $\min_{\boldsymbol{X}\in\mathcal{X}} f(\boldsymbol{X})$

$$
\begin{aligned}
\min_{\boldsymbol{Y}\in\mathcal{X},q} \; & q \\
\text{s.t.} \; & q \geq f(\boldsymbol{X}) + \langle d^f_{\boldsymbol{X}}(\boldsymbol{X}), (\boldsymbol{Y}-\boldsymbol{X})\rangle_F, \quad \forall \boldsymbol{X}\in\mathcal{X}
\end{aligned}
$$

subgradient of $f(\boldsymbol{X})$

Replaced by $\mathcal{X}_{Bender}$ (Benders decomposition)

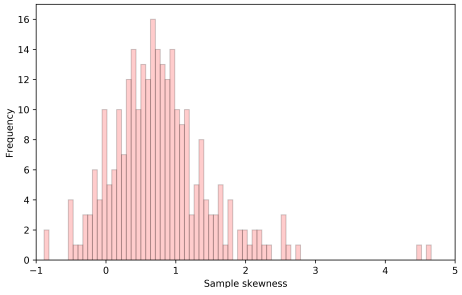
Why SVF Can Fail

Smallest variance first (SVF)

- Works under strong ordering conditions such as dilation order.
- Easy to implement and widely used as a sequencing heuristic.
- But **variance is symmetric**: it does not distinguish long right tails from long left tails.

Key observation

Under the "to follow" practice, a long right tail is especially costly because delays propagate forward.



Most surgery duration distributions are right-skewed in the data.

Two heuristics for C_α.

- **Standard deviation (SD)**: the induced sequence is exactly SVF.

$$
\begin{aligned}
C_\alpha(\tilde{t}) = &\ \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] + \alpha \ln \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{\tilde{t} - \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right]}{\alpha}\right)\right] \\
\approx &\ \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] + \alpha \ln \mathbb{E}_{\mathbb{P}}\left[1 + \frac{\tilde{t} - \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right]}{\alpha} + \frac{(\tilde{t} - \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right])^2}{2\alpha^2}\right] \\
\approx &\ \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] + \alpha \frac{\mathbb{E}_{\mathbb{P}}\left[\tilde{t} - \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right]\right]^2}{2\alpha^2} \\
= &\ \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] + \frac{\sigma(\tilde{t})^2}{2\alpha}.
\end{aligned}
$$

Two heuristics for C_α.

- **Forward and backward deviations (FBD)**
  Forward deviation (FD): $\xi(\tilde{t}) = \sqrt{\sup_{\alpha>0} 2\alpha C_\alpha(\tilde{t} - \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right])}$.
  Backward deviation (BD): $\zeta(\tilde{t}) = \sqrt{\sup_{\alpha>0} 2\alpha C_\alpha(\mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] - \tilde{t})}$.

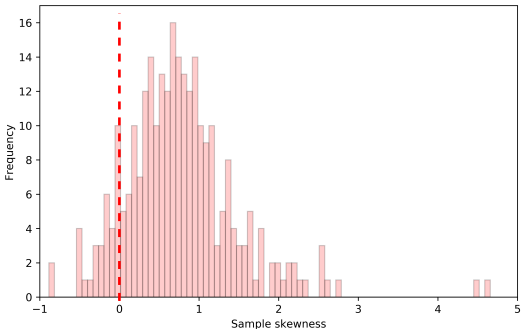
$$C_\alpha(\tilde{t}) \approx \mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] + \frac{\xi(\tilde{t})^2}{2\alpha} \qquad C_\alpha(-\tilde{t}) \approx -\mathbb{E}_{\mathbb{P}}\left[\tilde{t}\right] + \frac{\zeta(\tilde{t})^2}{2\alpha},$$

FBD properties

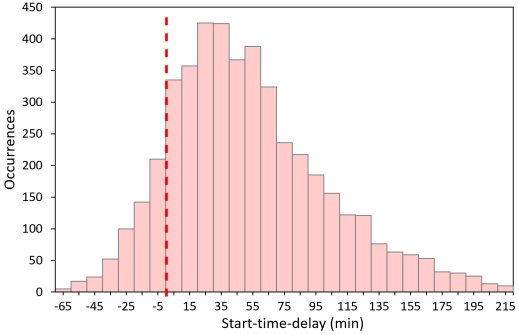
- If $\mathbb{E}_{\mathbb{P}}[\tilde{v} - \mathbb{E}_{\mathbb{P}}[\tilde{v}]]^3 > 0$, then $\xi(\tilde{v}) > \sigma(\tilde{v})$. Further, if $\mathbb{E}_{\mathbb{P}}[\tilde{v} - \mathbb{E}_{\mathbb{P}}[\tilde{v}]]^3 < 0$, then $\zeta(\tilde{v}) > \sigma(\tilde{v})$.
- If the dilation order exists, both FD and BD have the same order.

Data from a Singapore Hospital

- Data on surgeries conducted each day and their actual durations.
- 6 operating rooms; two years of data
- 1,177 types of surgeries and 10,294 elective surgeries.
- Among surgery types with more than 10 samples: 88.7% are right-skewed.



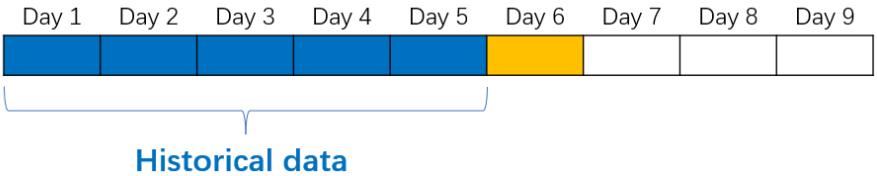
(a) Skewness



(b) Delay

Experimental Settings

- **Data collection**: rolling horizon.

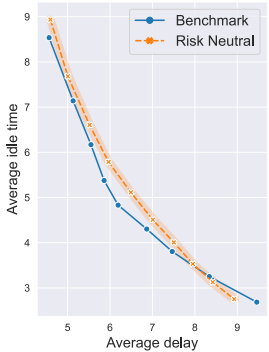


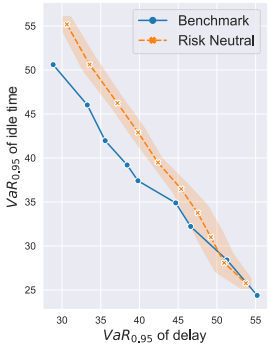
- **Effective instances**: all surgeries have more than 10 historical samples.
- **Tolerance levels**: $\tau_1 = 30, \tau_2 = 15$.
- **Evaluation criteria**: average, $VaR_{0.95}$, and $CVaR_{0.95}$ of delay $d = \max(a - s - \tau_1, 0)$ and idle time $i = \max(s - \tau_2 - a, 0)$, where a is the actual start time.

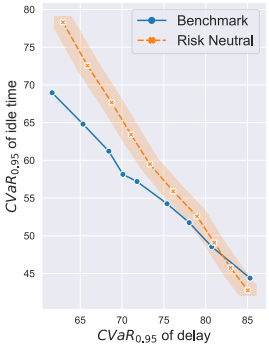
Benchmark and Risk-neutral Model

Benchmark: Use the empirical sample distribution to compute C_α.
Risk-neutral model:

$$\sum_{j\in[N-1]}\left(\omega_1\mathbb{E}\left[\left(\sum_{i\in[j]}\sum_{n\in[N]}x_{in}\tilde{t}_n-s_{j+1}-\tau_1\right)^+\right]+\omega_2\mathbb{E}\left[\left(s_{j+1}-\sum_{i\in[j]}\sum_{n\in[N]}x_{in}\tilde{t}_n-\tau_2\right)^+\right]\right).$$







Interpreting Risk-neutral Comparison

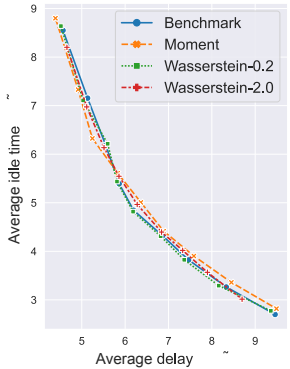
- The Punctuality Index controls tail-sensitive schedule risk, not just average delay or idle time.
- Risk-neutral scheduling may seem better if we only target averages.
- In high-delay-risk cases, the Punctuality Index improves trade-offs, especially for $VaR_{0.95}$ and $CVaR_{0.95}$.

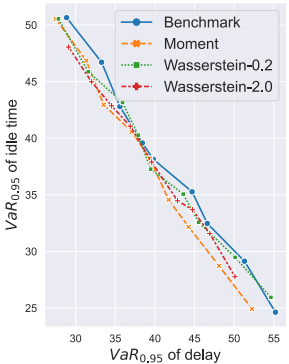
Takeaway

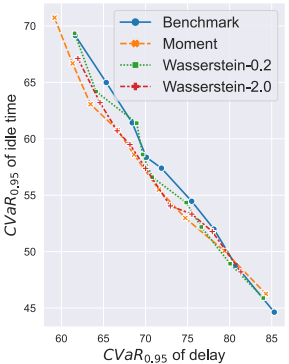
This measure is not just a preference; it moves schedules to better protect against severe deviations from the acceptable window.

Benchmark and DRO Approaches

Moment: DRO with ambiguity in the mean and mean absolute deviation.
Wasserstein-θ: DRO with Wasserstein ambiguity; θ is the Wasserstein radius.







Interpreting DRO Comparison

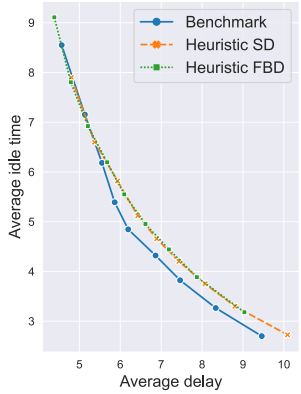
- The benchmark uses the empirical distribution.
- DRO enlarges the set of plausible duration distributions.
- This may be conservative for averages but reduces extreme delay and idle-time risk.

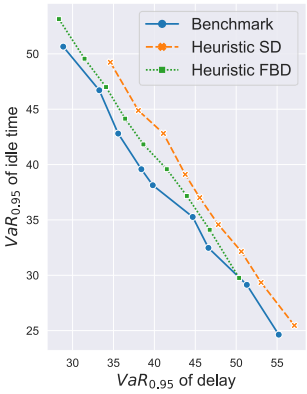
Managerial interpretation

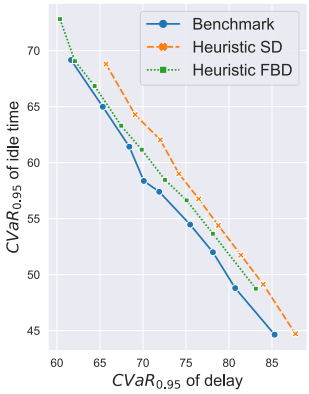
With limited samples, DRO asks: "Which schedules stay reliable if the empirical distribution is slightly misspecified?"

Benchmark and Heuristics

SD: Heuristic based on standard deviation.
FBD: Heuristic based on forward and backward deviations.







Interpreting Heuristics

- SD is simple and matches SVF when the normal approximation holds.
- FBD measures forward and backward deviations, so it handles skewed duration distributions.
- In our data, FBD is usually more robust than SD, with only a small loss versus the exact benchmark.

Practical message

If the exact mixed-integer convex model is too costly, FBD gives a simple sequencing rule that preserves the skewness information managers care about.

Takeaways

- **Punctuality Index**
  - Interpretation: evaluates the risks of delay and idling, takes both the violation probability and magnitude into account, and is risk-averse.
  - Tractability: additivity and joint convexity.
- Distributional flexibility: accommodates both **known distributions and distributional ambiguity sets**.
- Tractable formulation and solution method for both **sequencing and scheduling decisions**.
- Heuristics that consider the distributional **asymmetry**.

Reference I

Bai M, Storer RH, Tonkay GL, Theman TE (2020) Reinvestigating surgery scheduling in the "to-follow" practice. *Available at SSRN 3685538* .

Baker KR (2014) Minimizing earliness and tardiness costs in stochastic scheduling. *European Journal of Operational Research* 236(2):445–452.

Balzer C, Raackow D, Hahnenkamp K, Flessa S, Meissner K (2017) Timeliness of operating room case planning and time utilization: influence of first and to-follow cases. *Frontiers in Medicine* 4:49.

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* 35(11):1003–1016.

Denton B, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science* 10:13–24.

Dexter F, Epstein RH, Traub RD, Xiao Y, Warltier DC (2004) Making management decisions on the day of surgery based on operating room efficiency and patient waiting times. *Anesthesiology* 101(6):1444–1453.

Doebbeling BN, Burton MM, Wiebke EA, Miller S, Baxter L, Miller D, Alvarez J, Pekny J (2012) Optimizing perioperative decision making: improved information for clinical workflow planning. *AMIA Annual Symposium Proceedings* 154–163.

Freeman NK, Melouk SH, Mittenthal J (2016) A scenario-based approach for operating theater scheduling under uncertainty. *Manufacturing & Service Operations Management* 18(2):245–261.

Fügener A, Schiffels S, Kolisch R (2017) Overutilization and underutilization of operating rooms-insights from behavioral healthcare operations management. *Health Care Management Science* 20(1):115–128.

Reference II

Guda H, Dawande M, Janakiraman G, Jung KS (2016) Optimal policy for a stochastic scheduling problem with applications to surgical scheduling. *Production and Operations Management* 25(7):1194–1202.

Guerriero F, Guido R (2011) Operational research in the management of the operating theatre: a survey. *Health Care Management Science* 14:89–114.

Gupta D (2007) Surgical suites' operations management. *Production and Operations Management* 16(6):689–700.

Jiang R, Ryu M, Xu G (2019) Data-driven distributionally robust appointment scheduling over wasserstein balls. *arXiv preprint arXiv:1907.03219* .

Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* 61(3):711–726.

Mak HY, Rong Y, Zhang J (2014) Appointment scheduling with limited distributional information. *Management Science* 61(2):316–334.

Macario A (2010) What does one minute of operating room time cost? *Journal of Clinical Anesthesia* 22(4):233–236.

Mancilla C, Storer RH (2012) Stochastic integer programming based algorithms for adaptable open block surgery scheduling.

Otten M, Braaksma A, Boucherie RJ (2019) Minimizing earliness/tardiness costs on multiple machines with an application to surgery scheduling. *Operations research for health care* 22:100194.

Reference III

Pinedo ML (2012) *Scheduling*, volume 29 (Springer).

Qi J (2016) Mitigating Delays and Unfairness in Appointment Systems. *Management Science* 63(2):566-583.

Sauré A, Begen MA, Patrick J (2020) Dynamic multi-priority, multi-class patient scheduling with stochastic service times. *European Journal of Operational Research* 280(1):254–265.

Shylo OV, Prokopyev OA, Schaefer AJ (2012) Stochastic operating room scheduling for high-volume specialties under block booking. *INFORMS Journal on Computing* 25(4):682–692.

van Essen JT, Hurink JL, Hartholt W, van den Akker BJ (2012) Decision support system for the operating room rescheduling problem. *Health Care Management Science* 15:355–372.

Wachtel RE, Dexter F (2009) Tactical increases in operating room block time for capacity planning should not be based on utilization. *Anesthesia & Analgesia* 108(1):215–226.

Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production and Operations Management* 23(5):788–801.

Zhang Y, Wang Y, Tang J, Lim A (2020) Mitigating overtime risk in tactical surgical scheduling. *Omega* 93:102024.