Moment-based Distributionally Robust Server Allocation and Scheduling Problems

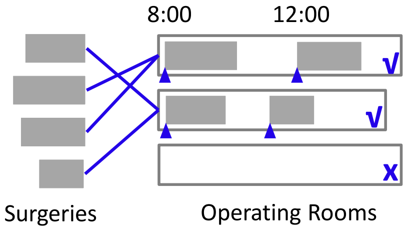
Yiling Zhang[1], Siqian Shen[1], Ayca Erdogan[2]

[1]: Dept. of IOE, University of Michigan
[2]:Dept. of ISE, San José State University

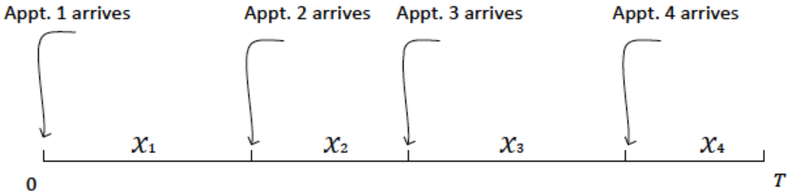
Outline

Two Common Problems in Service Operations

P1: Server Allocation



P2: Appointment Scheduling



Generic Problem Settings

Common issues: 1) service time uncertainty; 2) unknown distributions with limited data.

Allocation phase: Given a set of servers and jobs:

- Decisions: Which servers to open and how to allocate jobs.
- Objective: Minimize the total operational cost.
- Constraint: Low overtime probability in each open server.

Generic Problem Settings

Common issues: 1) service time uncertainty; 2) unknown distributions with limited data.

Allocation phase: Given a set of servers and jobs:

- Decisions: Which servers to open and how to allocate jobs.
- Objective: Minimize the total operational cost.
- Constraint: Low overtime probability in each open server.

Scheduling phase: Given appointments assigned to a server:

- Decisions: Arrival time of each appointment
- Objective: Minimize the total waiting (+ idleness)
- Constraint: Low overtime probability

Literature Review

Allocation:

- Deterministic: Blake and Donald (2002), Jebali et al. (2006)
- Stochastic multi-OR allocation: Denton et al. (2010)
- Chance-constrained multi-OR allocation: Shylo et al. (2012)

Scheduling:

- Under random service durations: Weiss (1990), Van den Bosch and Dietz (2000), Denton and Gupta (2003), Gupta and Denton (2008), Pinedo (2012), Erdogan and Denton (2013)
- Near-optimal scheduling policy: Mittal et al. (2014), Begen and Queyranne (2011), Begen et al. (2012), Ge et al. (2013)
- Simulation and queuing theories: Bailey (1952); Brahimi and Worthington (1991); Ho and Lau (1992); Rohleder and Klassen (2002); Hassin and Mendel (2008); Zeng et al. (2010)
- Distributionally Robust (DR) appointment scheduling: Mak et al. (2014) and Kong et al. (2014)

In this Talk...

Under random service time, we consider

- Problem 1: Multiple Server Allocation;
- Problem 2: Single Server Appointment Scheduling

We study their Distributionally Robust (DR) variants, and employ

- Moment ambiguity sets of the unknown distribution

We reformulate the DR models as

- Allocation: 0-1 SDP (cross-moment), 0-1 SOCP (exact 1st & 2nd-moment matching), 0-1 SOCP (Gaussian Approximation)
- Scheduling: SDP (cross-moment ambiguity set)

We optimize the 0-1 SDP via a cutting-plane algorithm, and directly compute the rest in off-the-shelf solvers.

Outline

Notation

- Set of Servers: I (operating cost τ_i and time limit T_i)
- Set of Jobs: J ($\rho_{ij} = 1$ if job j can be operated on server i)
- Random service durations: $s = [s_{ij},\ i \in I,\ j \in J]^{\mathsf{T}}$
- Decision Variable
  - $z_i \in \{0, 1\}$: whether or not to operate server i, such that

$$z_i = \begin{cases} 1 & \text{operate server } i \\ 0 & \text{o.w.} \end{cases}$$

  - $y_{ij} \in \{0, 1\}$: whether to assign job j to server i, with

$$y_{ij} = \begin{cases} 1 & \text{allocate job } j \text{ to server } i \\ 0 & \text{o.w.} \end{cases}$$

0-1 Chance-Constrained Formulation

Let α_i be the risk tolerance of having overtime on server i, $\forall i \in I$.

$$\begin{aligned}
\min_{\mathbf{z},\mathbf{y}} \quad & \sum_{i \in I} \tau_i z_i \\
\text{s.t.} \quad & y_{ij} \le \rho_{ij} z_i, \ \forall i \in I, \ j \in J \\
& \sum_{i \in I(j)} y_{ij} = 1, \ \forall j \in J \\
& \mathbb{P}\left\{ \sum_{j \in J(i)} s_{ij} y_{ij} \le T_i \right\} \ge 1 - \alpha_i, \ \forall i \in I \\
& y_{ij}, z_i \in \{0, 1\}, \ \forall i \in I, \ j \in J.
\end{aligned}$$

A variant of chance-constrained binary packing (see, e.g., Song, Luedtke, and Küçükyavuz (2014))

Moment-based Ambiguity Sets

Consider $s_i = [s_{ij},\ j \in J]^\mathsf{T}$ as random service time of server i. Due to limited data, we may not know the exact distributions of s_i, and thus cannot accurately evaluate $\mathbb{P}\left\{\sum_{j \in J(i)} s_{ij} y_{ij} \leq T_i\right\}$. Thus, we consider

- Cross-moment Ambiguity Set (Delage and Ye (2010)):

$$\mathcal{D}_M^i(\mu_0^i, \Sigma_0^i, \gamma_1, \gamma_2) = \left\{ f(s_i) : \begin{array}{l} \int_{s_i \in \Xi_i} f(s_i) ds_i = 1 \\ (\mathbb{E}[s_i] - \mu_0^i)^\mathsf{T} (\Sigma_0^i)^{-1} (\mathbb{E}[s_i] - \mu_0^i) \leq \gamma_1 \\ \mathbb{E}[(s_i - \mu_0^i)(s_i - \mu_0^i)^\mathsf{T}] \preceq \gamma_2 \Sigma_0^i \end{array} \right\}$$

- Special Case Ambiguity Set (Exact Mean and Covariance Matching):

$$\mathcal{D}_C^i(\mu_0^i, \Sigma_0^i) = \left\{ f(s_i) : \quad \begin{array}{l} \int_{s_i \in \Xi_i} f(s_i) ds_i = 1,\ \mathbb{E}[s_i] = \mu_0^i \\ \mathbb{E}[(s_i - \mu_0^i)(s_i - \mu_0^i)^\mathsf{T}] = \Sigma_0^i \end{array} \right\}$$

DR Chance Constraint

- A DR Allocation Model: Replace

$$\mathbb{P}\left\{\sum_{j\in J(i)} s_{ij}y_{ij} \leq T_i\right\} \geq 1-\alpha_i,\ \forall i \in I$$

with

$$\inf_{f(s_i)\in\mathcal{D}} \mathbb{P}\left\{\sum_{j\in J} s_{ij}y_{ij} \leq T_i\right\} \geq 1-\alpha_i,\ \forall i \in I.$$

where $\mathcal{D}$ is either $\mathcal{D}_M^i$ or $\mathcal{D}_C^i$.

Outline

Allocation $\Rightarrow$ 0-1 SDP when $\mathcal{D} = \mathcal{D}_M^i$

To reformulate $\inf_{f(s_i)\in\mathcal{D}} \mathbb{P}\left\{\sum_{j\in J} s_{ij} y_{ij} \leq T_i\right\} \geq 1-\alpha_i$, define

- $\begin{bmatrix} H^i & p^i \\ (p^i)^\mathsf{T} & q^i \end{bmatrix}$: dual of $(\mathbb{E}[s_i]-\mu_0^i)^\mathsf{T}(\Sigma_0^i)^{-1}(\mathbb{E}[s_i]-\mu_0^i) \leq \gamma_1$
- G^i: dual variables with $\mathbb{E}[(s_i-\mu_0^i)(s_i-\mu_0^i)^\mathsf{T}] \preceq \gamma_2 \Sigma_0^i$
- r^i: dual variables with $\int_{s_i\in\Xi_i} f(s_i)ds_i = 1$.

Following Jiang and Guan (2015),

- the DR chance constraint is equivalent to SDP constraints.
- the DR server allocation model then becomes a 0-1 SDP.

Thus, we propose a cutting-plane algorithm that decomposes the 0-1 SDP into two stages.

Master Problem: 0-1 Integer Linear Program

A Master Problem (MP) without enforced DR chance constraints:

$$\begin{aligned}
\min_{\mathbf{z},\mathbf{y}} \quad & \sum_{i\in I} \tau_i z_i \\
\text{s.t.} \quad & y_{ij} \leq \rho_{ij} z_i, \ \forall i \in I, \ j \in J \\
& \sum_{i\in I(j)} y_{ij} = 1, \ \forall j \in J \\
& \mathcal{C}_i(y_i) \leq 0, \ i \in I \\
& y_{ij}, z_i \in \{0,1\}, \ \forall i \in I, \ j \in J,
\end{aligned}$$

where $\mathcal{C}_i(y_i) \leq 0$ include linear cuts from solving server-based subproblems that evaluate whether y can satisfy the server-based DR chance constraints.

Subproblem Dual and Valid Cuts

Given y from MP, we formulate a subproblem for each $i \in I$ as the equivalent SDP of the DR chance constraint by letting $\mathcal{D} = \mathcal{D}^i_M$.

Take the dual of the SDP subproblem (also an SDP):

$$\text{SUB}^i(y_i)\text{-Dual: } \max_{Q^i, d^i, u^i} \quad y_i^{\mathsf{T}} d^i + (y_i^{\mathsf{T}} \mu_0^i - T_i) u^i \le 0$$

$$\begin{bmatrix} \gamma_2 \Sigma_0^i & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} Q^i & d^i \\ (d^i)^{\mathsf{T}} & u^i \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} 0 & 0 \\ 0 & -\alpha_i \end{bmatrix} + \begin{bmatrix} Q^i & d^i \\ (d^i)^{\mathsf{T}} & u^i \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} Q^i & d^i \\ (d^i)^{\mathsf{T}} & u^i \end{bmatrix} \in S_+^{(|J(i)|+1)\times(|J(i)|+1)}.$$

Consider optimal $(\tilde{d}^i, \tilde{u}^i)$. If $y_i^{\mathsf{T}} \tilde{d}^i + (y_i^{\mathsf{T}} \mu_0^i - T_i)\tilde{u}^i > 0$, then generate a valid cut (linear in y_i).

A Cutting-Plane Approach

1. Initial MP without $\mathcal{C}_i(y_i) \leq 0, \; i \in I$.
2. Iterate the following steps until no cuts are needed:
   i. Solve MP and obtain (z, y). If fail, claim infeasible, exit.
   ii. Otherwise, for $i \in I$ do
      - Solve $\text{SUB}^i(y_i)$-Dual and obtain optimal dual (Q^i, d^i, u^i).
      - If $((d^i)^\mathsf{T} + d^i(\mu_0^i)^\mathsf{T})y_i - u^i T_i > 0$, generate a cut

        $$((d^i)^\mathsf{T} + u^i(\mu_0^i)^\mathsf{T})y_i - u^i T_i \leq 0$$

        into cut set $\mathcal{C}_i(y_i) \leq 0$ of MP.
   iii. If no cut generated from $\text{SUB}^i(y_i)$-Dual for $\forall i \in I$, then (z, y) is optimal; exit.

Allocation $\Rightarrow$ 0-1 SOCP when $\mathcal{D} = \mathcal{D}_C^i$

We replace $\inf_{f(s_i)\in\mathcal{D}} \mathbb{P}\left\{\sum_{j\in J} s_{ij}y_{ij} \leq T_i\right\} \geq 1-\alpha_i$ by an SOCP constraint given:

Theorem (Wagner, 2008)

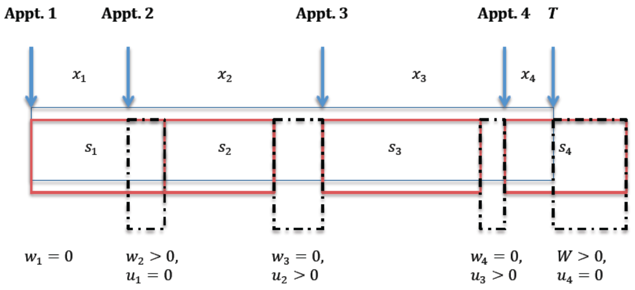
Given the first and second order information μ_0^i and Σ_0^i of the service duration vector s_i, given the ambiguity set $\mathcal{D}_C^i$ and probability α_i, then an equivalent formulation for $\inf_{f(s_i)\in\mathcal{D}_C^i} \mathbb{P}[s_i^\mathsf{T} y_i \leq T_i] \geq 1-\alpha_i$ is

$$\sqrt{y_i^\mathsf{T}\Sigma_0^i y_i} \leq \sqrt{\frac{\alpha_i}{1-\alpha_i}}(T_i - (\mu_0^i)^\mathsf{T} y_i), \ \forall i \in I.$$

Alternatively, the DR allocation model is a 0-1 SOCP and is directly optimized by CVX 2.1 + Gurobi solver.

Outline

Appointment Scheduling: Notation



Parameters:

- One server and m appt. arriving in a fixed order
- Service durations: s_j
- Unit waiting penalty: h_j

Decision variables:

- x_j: time interval between appt. j and $j+1$.
- w_j: waiting time of appt. j

Scheduling: Chance-Constrained Linear Program

$$
\begin{aligned}
\min_{x} \quad & \mathbb{E}_{s:f(s)}\left[\min_{w} \sum_{j=2}^{m} h_j w_j\right] \\
\text{s.t.} \quad & \mathbb{P}\left\{\sum_{j=1}^{m-1} x_j + w_m + s_m \leq T\right\} \geq 1 - \alpha \\
& w_j + x_{j-1} \geq s_{j-1} + w_{j-1},\ \forall j = 2, \ldots, m \\
& x_j \geq 0, \quad \forall j = 1, \ldots, m-1 \\
& w_1 = 0,\ w_j \geq 0,\ \forall j = 2, \ldots, m,
\end{aligned}
$$

- Balance waiting of appointments and server overtime.
- Remain the same complexity if adding idle-time penalty.

A DR Variant

We employ the cross-moment ambiguity set

$$\mathcal{D}_M^s = \left\{ f(s): \begin{array}{l} \int_{s\in\Xi^s} f(s)ds = 1,\ (\mathbb{E}[s]-\mu_0^s)^{\mathsf{T}}(\Sigma_0^s)^{-1}(\mathbb{E}[s]-\mu_0^s) \le \gamma_1 \\ \mathbb{E}[(s-\mu_0^s)(s-\mu_0^s)^{\mathsf{T}}] \preceq \gamma_2 \Sigma_0^s \end{array} \right\}.$$

- Worst Case Expected Waiting Penalty:

$$\min_x \max_{f(s)\in\mathcal{D}_M^s} \mathbb{E}_{f(s)} \left[\min_w \sum_{j=2}^{m} h_j w_j \right]$$

- DR Chance Constraint on Overtime:

$$\inf_{f(s)\in\mathcal{D}_M^s} \mathbb{P}\left\{ \sum_{j=1}^{m-1} x_j + w_m + s_m \le T \right\} \ge 1-\alpha$$

Reformulation: Key Ideas

Following similar procedures in the DR allocation:

- DR Chance Constraint $\Rightarrow$ multiple # of SDP
- Worst Case Expectation $\Rightarrow$ semi-infinite SDP with infinite # of constraints
- Use the extreme-point representation of the dual of the linear scheduling constraints (special structure in Mak et al. (2014))
- Reformulate the SDP with semi-infinite constraints as SDP

The overall DR scheduling problem with cross-moment ambiguity set is an SDP and optimized directly in CVX 2.1 + Gurobi.

Outline

Allocation Setup

Gaussian distributed $s_{ij} \Rightarrow$ a benchmark 0-1 SOCP model.
Solver: Matlab-based CVX 2.1 + gurobi solver

Experimental setup:

- 32 jobs, 6 servers
- Each server: time limit = 8 hrs, operating cost = 1.
- 4 combinations of
  - High mean (20min–30min) or Low mean (10min–15min)
  - High variance (CoV = 1) or Low variance (CoV = 0.3)
- 5 sets of tests:
  - `eq`: 32 jobs with equally mixed types; 8 each.
  - `ll`, `lh`, `hl`, `hh`: a certain type of jobs dominate. (The first letter refers to "mean" and the second refers to "variance").
- Training samples follow Gamma distributions
- Training data size = 20 for each type

Average CPU Time

We report the CPU seconds for computing each type instance with different methods by letting $\alpha = 0.05$ and $\alpha = 0.10$.

α	Approach	eq	ll	hl	lh	hh
0.05	Gaussian	1.62	1.78	1.70	1.59	170.68
	0-1 SOCP	23.56	6.22	57.10	6.68	1096.92
	Cutting-Plane	47.41	29.78	49.76	30.61	233.22
0.10	Gaussian	1.65	1.79	1.78	1.34	2.15
	0-1 SOCP	14.76	7.85	8.72	7.46	18.42
	Cutting-Plane	23.96	33.20	45.10	28.44	174.85

Solution Performance

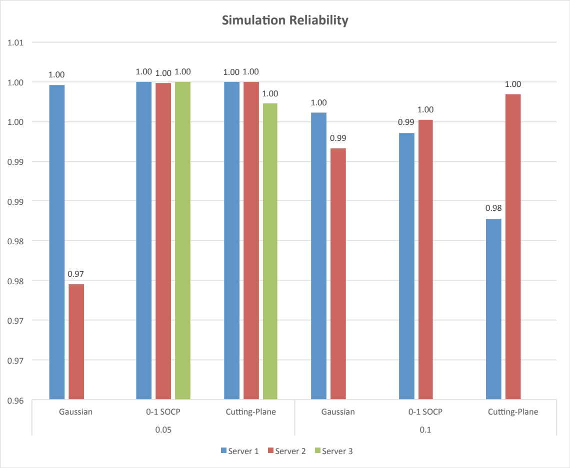
Table: # of servers opened by each method

α	Gaussian	0-1 SOCP	Cutting-Plane
0.05	2	3	3
0.1	2	2	2

Taking the setting eq:

- Follow "Lognormal" to generate $10,000$ data for simulation.
- Fix solutions to the three models in the simulation sample and evaluate how many scenarios are satisfied.
- Report the results of "training sample" $=$ gamma, and "simulation sample" (i.e., true distribution) $=$ lognormal.

Probability of No Overtime in Simulation Sample



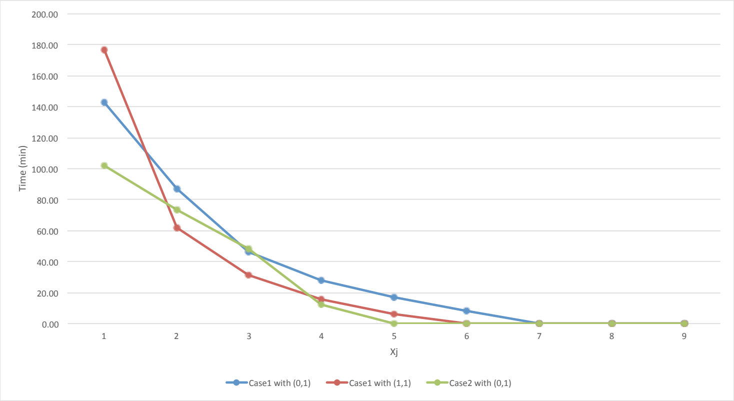
- Both 0-1 SDP and 0-1 SOCP provide highly reliable DR solutions.
- The opt. solution of the benchmark model based on Gaussian approximation performs slightly worse on Server #2.
- The performance is not sensitive to distribution change.

Outline

Scheduling Setup

- 10 appointments, 1 server (can be a DR allocation solution)
- Server time limit: 8 hours
- Unit waiting penalty with all appointments
- Tolerable overtime risk $\alpha = 0.1$
- Appointments arrive in the following two orders
  - Order 1: 4 `hh` $\rightarrow$ 3 `hl` $\rightarrow$ 3 `ll` appointments
  - Order 2: 3 `ll` $\rightarrow$ 3 `hl` $\rightarrow$ 4 `hh` appointments

Solution Pattern



- A more robust model intend to increase the time interval in between the first two appointments.
- As more `11` appointments appear at the beginning, we intend to distribute time intervals more evenly.

Waiting Time and Overtime 99% Quantiles

Table: 99 % quantiles of waiting and overtime (in min)

Appt.	Waiting (min)	eq	ll	hh	lh	hl
1 (hh)	w_1	0.00	0.00	0.00	0.00	0.00
2 (hh)	w_2	0.00	0.00	0.00	0.00	0.00
3 (hh)	w_3	35.68	0.00	37.65	0.00	0.00
4 (hh)	w_4	74.19	0.00	78.35	1.31	14.89
5 (hl)	w_5	99.60	0.00	92.18	18.86	30.13
6 (hl)	w_6	30.43	7.03	107.82	31.26	44.08
7 (hl)	w_7	39.20	15.76	117.83	38.93	50.65
8 (ll)	w_8	46.81	23.51	120.11	47.98	62.96
9 (ll)	w_9	23.34	23.92	124.24	47.92	60.03
10 (ll)	w_{10}	23.77	23.65	119.23	47.22	58.46
	Overtime (min)	0.00	0.00	22.73	0.00	0.00

Recall that the total time = 480 min.

Conclusions

Conclusions:

- Consider DR server allocation and DR appointment scheduling models and algorithms.
- Employ diverse moment-based ambiguity sets of distributions $\Rightarrow$ 0-1 SDP / 0-1 SOCP for allocation and SDP for scheduling.
- Develop cutting-plane algorithm for 0-1 SDP.

Future Research:

- Investigate other ambiguity sets.
- Study data-driven aspects of different sets.
- Implement in practice.