Optimal control of an emergency room triage and treatment process

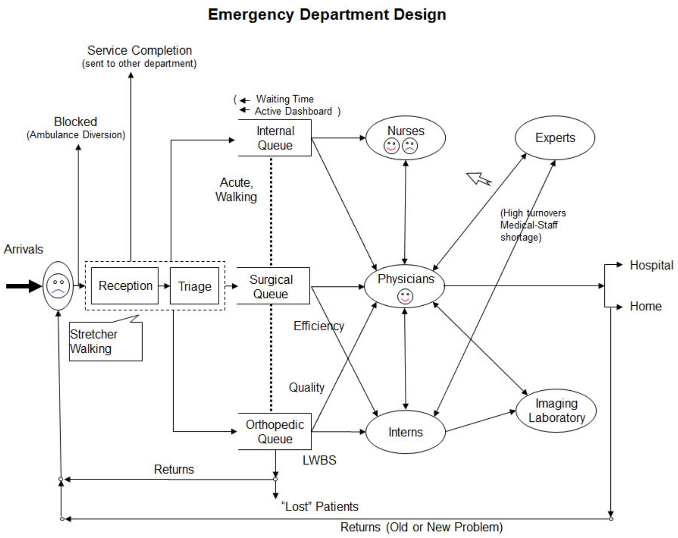
Gabriel Zayas-Cabán[1] Mark E. Lewis[1] Jungui Xie[2] Linda V. Green[3]

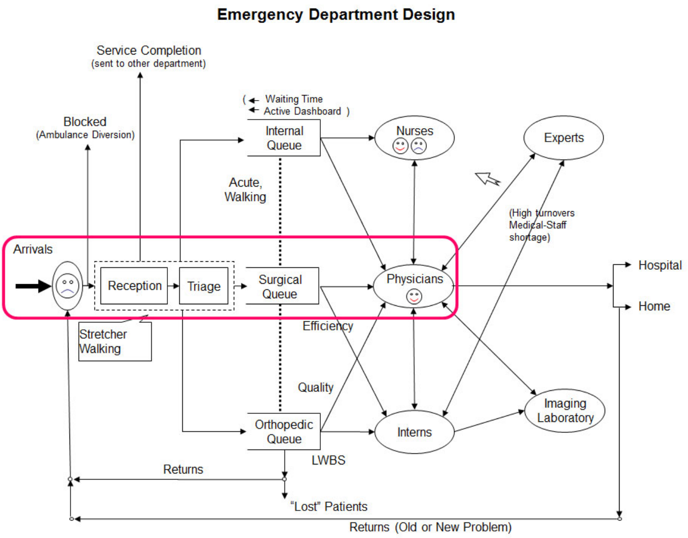
[1]Cornell University
Ithaca, NY

[2]University of Science and Technology of China
Beijing, China

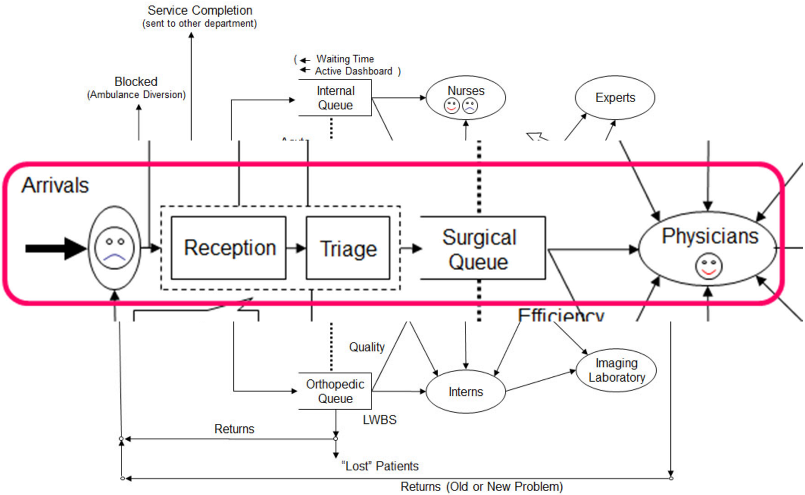
[3]Columbia University
New York, NY

Outline


Emergency Department Design
Service Completion
(sent to other department)
Waiting Time
Active Dashboard
Blocked
(Ambulance Diversion)
Internal Queue
Nurses
Experts
Acute, Walking
(High turnovers Medical-Staff shortage)
Arrivals
Reception
Triage
Surgical Queue
Physicians
Hospital
Home
Stretcher Walking
Efficiency
Quality
Imaging Laboratory
Orthopedic Queue
Interns
LWBS
Returns
"Lost" Patients
Returns (Old or New Problem)


Emergency Department Design
Service Completion
(sent to other department)
Blocked
(Ambulance Diversion)
(Waiting Time
Active Dashboard)
Internal Queue
Nurses
Experts
Acute, Walking
(High turnovers
Medical-Staff
shortage)
Arrivals
Reception
Triage
Surgical Queue
Physicians
Hospital
Home
Stretcher
Walking
Efficiency
Quality
Orthopedic Queue
Interns
Imaging Laboratory
LWBS
Returns
"Lost" Patients
Returns (Old or New Problem)

Emergency Department Design



NEW CARE MODELS IN THE ED

Emergency Department (ED):

- In 2010, number of visits in the U.S. around 129.8 million and increasing 2–3% per year.
- Number of ED beds decreasing.
- Overcrowded departments, long waiting times, overworked staff, patient dissatisfaction, and abandonments (LWOT). [NAMCS]

NEW CARE MODELS IN THE ED

- Many ED patients present with low-acuity conditions and do not require hospitalization.
- Low-acuity ED patients have to be treated, diverting resources from more critical patients.
- EDs developing new models of care to handle these lower-acuity patients to facilitate patient flow. [Helm *et al.* 2011, Saghafian *et al.* 2012, Saghafian *et al.* 2014]

The Lutheran Medical Center

Overview

Image available at http://www.lutheranmedicalcenter.com; downloaded June 2013.

Lutheran Medical Center (LMC) Triage-Treat-and-Release (TTR) program:

- Developed in 2010.
- Multiple providers (physicians or physician assistants) who handle both phases of service.

The Lutheran Medical Center

TTR program

1. Patients arrive to ED and are registered.
2. Patients proceed to triage (phase-one service) on a first-come-first-served (FCFS) basis.
3. After triage, high severity patients are assigned to another part of the ED for testing and/or treatment.
4. Low severity and low complexity patients await treatment (phase-two service) in triage area.

The Lutheran Medical Center

TTR program

- May help reduce long waiting times in the ER.
  - Earlier patient contact with a physician and, hence, earlier decision-making.
- Physicians and physician assistants are more reliable in assessing patients during triage. [Soremkun *et al.* 2012, Burströ *et al.* 2012]
- Decoupled (“Fast-track system”) vs. coupled (TTR Program) triage and treatment.
- Other examples: Health clinics, other ER operations.

Interested in

- **two-phase stochastic service systems**,
- having **single** medical service provider, and
- where patients may **renege** or **abandon** before completing service.

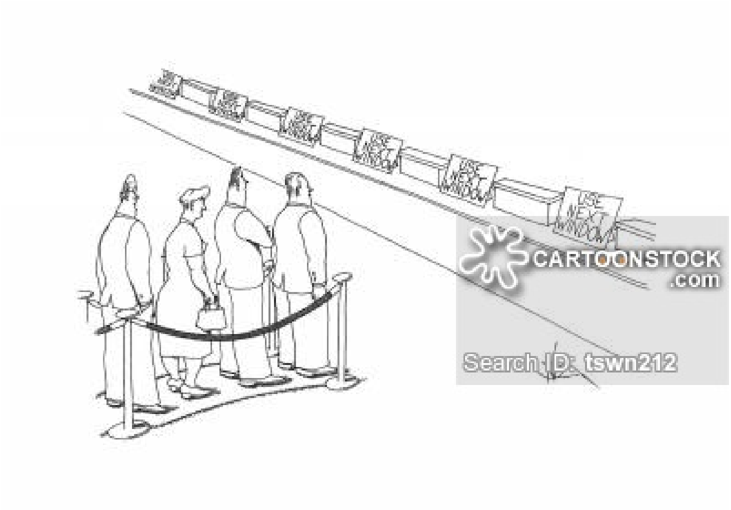
Interested in

- **two-phase stochastic service systems**,
- having **single** medical service provider, and
- where patients may **renege** or **abandon** before completing service.

Broad issue

How should we prioritize the work by medical service providers to balance initial delay for care with the need to discharge patients in a timely fashion.

A PRIMER ON QUEUEING SYSTEM



A Primer on Queueing System

Queueing system,

- One or more servers (physicians, physician assistants) providing service to arriving customers (patients).
- If all servers busy, customer (patient) join one or more queues (or lines) in front of servers, hence the name.
- Three components: **arrival process**, **service mechanism**, and **queue discipline**.

A PRIMER ON QUEUEING SYSTEM

Queueing System,

- **Arrival process:** how customers arrive to the system.
    - A_i — interarrival time between customer $i-1$ and i.
    - $\lambda = \frac{1}{\mathbb{E}(A_i)}$:= the arrival rate.

A Primer on Queueing System

Queueing System,

- **Arrival process:** how customers arrive to the system.
  - A_i — interarrival time between customer $i-1$ and i.
  - $\lambda = \frac{1}{\mathbb{E}(A_i)} :=$ the arrival rate.
- **Service mechanism:** how many servers, how are they organized.
  - S_i — service time of the ith arriving customer.
  - $\mu = \frac{1}{\mathbb{E}(S_i)} :=$ the service rate.

A Primer on Queueing System

Queueing System,

- **Arrival process:** how customers arrive to the system.
  - A_i — interarrival time between customer $i-1$ and i.
  - $\lambda = \frac{1}{\mathbb{E}(A_i)}$:= the arrival rate.
- **Service mechanism:** how many servers, how are they organized.
  - S_i — service time of the ith arriving customer.
  - $\mu = \frac{1}{\mathbb{E}(S_i)}$:= the service rate.
- **Queue discipline:** rule used to choose next customer from queue when server completes service of current customer (e.g. FCFS).

Typically,

- Fix queueing system/model configuration.
- Use model to help evaluate and predict performance of existing and proposed system (e.g. waiting times, queue length, utilization).

Typically,

- Fix queueing system/model configuration.
- Use model to help evaluate and predict performance of existing and proposed system (e.g. waiting times, queue length, utilization).
- Theory and/or simulation experimentation.
- **Goal:** Improve the design of a system.

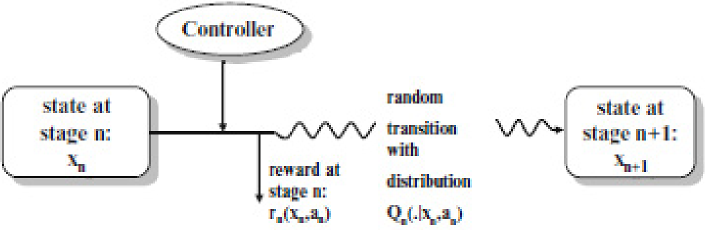
However,

- The parameters of the system (e.g. the arrival and service rates, queue disciplines) can be varied dynamically over time.
- Can significantly improve performance (e.g. reduced congestion, time spent waiting to be served).

However,

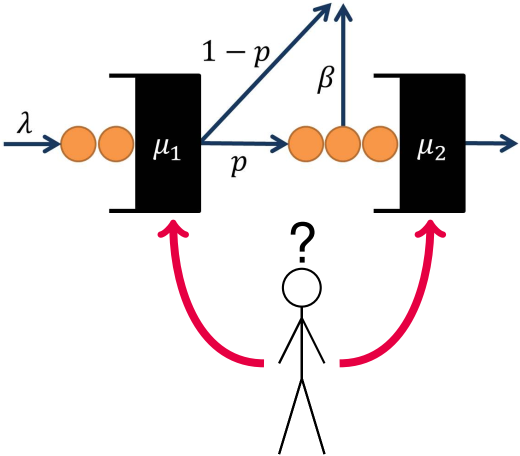
- The parameters of the system (e.g. the arrival and service rates, queue disciplines) can be varied dynamically over time.
- Can significantly improve performance (e.g. reduced congestion, time spent waiting to be served).
- **Markov decision processes.** [M. Puterman 2005]

Markov Decision Process Primer

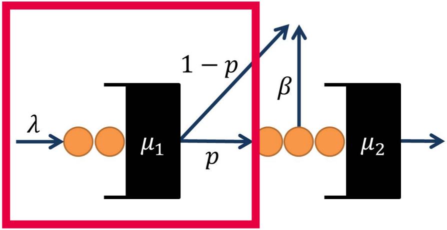


[Bäurle and Rieder, Markov Decision Processes with Applications to Finance]

Single-server tandem queue:

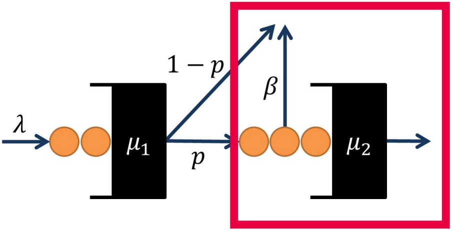


Single-server two-phase stochastic service system model:



- Rate λ Poisson arrival process.
- FCFS phase-one service (triage).
- After phase-one:
  - patients leave the system (w/ probability $1-p$), or
  - patients wait for FCFS phase-two service (w/ probability p).
  - $0 \leq p \leq 1$.

Single-server two-phase stochastic service system model:



- Patients wait for phase-two service (treatment) according to an exponentially distributed random variable with rate β before abandoning.
- Services in both phases are exponential with rates μ_1 and μ_2.
- After phase-two service, patient leaves the system.

Decision-making scenario:

1. Decision-maker (medical service provider) views number of patients at each station.

Decision-making scenario:

1. Decision-maker (medical service provider) views number of patients at each station.
2. Decides where to serve next, assuming **preemptive** service disciplines and rewards R_1 and R_2.

Decision-making scenario:

1. Decision-maker (medical service provider) views number of patients at each station.
2. Decides where to serve next, assuming **preemptive** service disciplines and rewards R_1 and R_2.

Specific objective

Want service disciplines that maximize total discounted expected reward or long-run average reward of the system.

State Space:

$$\mathbb{X} := \{(i,j) | i,j \in \mathbb{Z}^+\},$$

where i (j) represents number of patients at station 1 (2).

Decision epochs:

$$T := \{t_n, n \geq 1\},$$

sequence of times of events.

State Space:

$$\mathbb{X} := \{(i,j) | i,j \in \mathbb{Z}^+\},$$

where i (j) represents number of patients at station 1 (2).

Decision epochs:

$$T := \{t_n, n \geq 1\},$$

sequence of times of events.

Available actions in state $x = (i,j)$:

$$A(x) = \begin{cases} \{0,1,2\} & \text{if } i,j \geq 1, \\ \{0,1\} & \text{if } i \geq 1, j = 0, \\ \{0,2\} & \text{if } j \geq 1, i = 0, \\ \{0\} & \text{if } i = j = 0, \end{cases}$$

where $0, 1$, and 2 denote idling, serving at station 1, and serving at station 2.

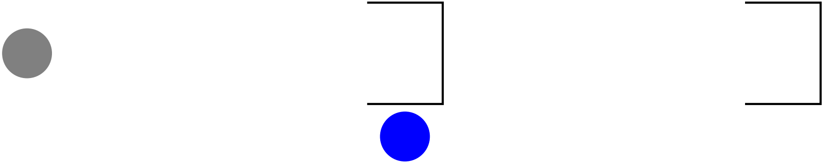
Reward: R_i received after completing phase i service, $i = 1, 2$.

Expected reward function:

$$r((i,j),a) = \begin{cases} \frac{\mu_1 R_1}{\lambda+\mu_1+j\beta} & \text{if } i > 0,\ a = 1, \\ \frac{\mu_2 R_2}{\lambda+\mu_2+j\beta} & \text{if } j > 0,\ a = 2, \\ 0 & \text{if } a = 0. \end{cases}$$

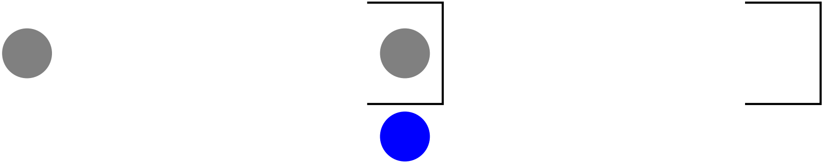
PRIORITIZE STATION 2 (P2)

PRIORITIZE STATION 2 (P2)



Prioritize Station 2 (P2)

Prioritize Station 2 (P2)

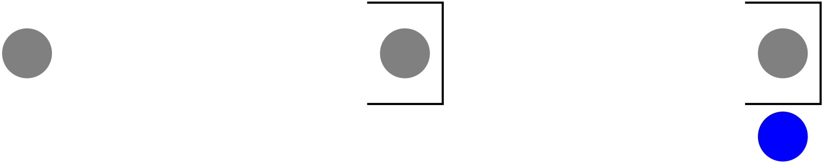


PRIORITIZE STATION 2 (P2)

PRIORITIZE STATION 2 (P2)

PRIORITIZE STATION 2 (P2)

Prioritize Station 2 (P2)



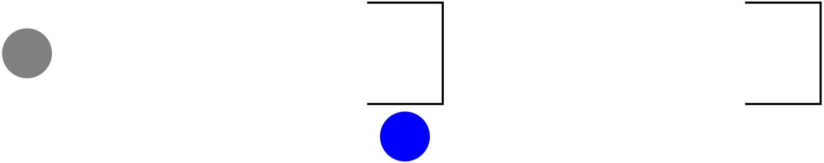
PRIORITIZE STATION 2 (P2)

PRIORITIZE STATION 2 (P2)

PRIORITIZE STATION 2 (P2)

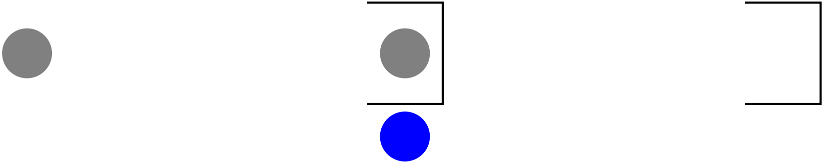
Prioritize Station 1 (P1)

Prioritize Station 1 (P1)



PRIORITIZE STATION 1 (P1)

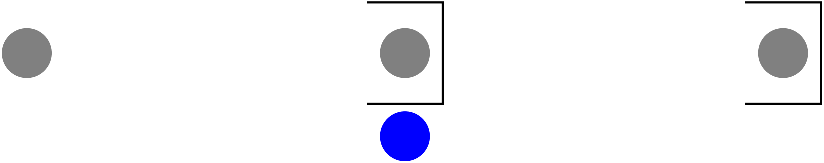
PRIORITIZE STATION 1 (P1)



Prioritize Station 1 (P1)

Prioritize Station 1 (P1)

PRIORITIZE STATION 1 (P1)



PRIORITIZE STATION 1 (P1)

PRIORITIZE STATION 1 (P1)

PRIORITIZE STATION 1 (P1)

PRIORITIZE STATION 1 (P1)

Prioritize Station 1 (P1)

Some results

Proposition

There is an optimal policy which does not idle the server whenever there are patients waiting.

SOME RESULTS

Proposition

There is an optimal policy which does not idle the server whenever there are patients waiting.

Theorem

The following hold:

1. *If* $\mu_2 R_2 \geq \mu_1 R_1$ *implies it is optimal to prioritize station 2.*
2. *If* $\lambda\left(\frac{1}{\mu_1} + \frac{1}{\mu_2+\beta}\right) < 1$ *and there is no discounting, then it is optimal to prioritize station 2.*

Some results

Proposition

There is an optimal policy which does not idle the server whenever there are patients waiting.

Theorem

The following hold:

1. *If* $\mu_2 R_2 \geq \mu_1 R_1$ *implies it is optimal to prioritize station 2.*
2. *If* $\lambda\left(\frac{1}{\mu_1} + \frac{1}{\mu_2+\beta}\right) < 1$ *and there is no discounting, then it is optimal to prioritize station 2.*

Proposition

If patients do not abandon, then $\mu_1 R_1 > \mu_2 R_2$ *implies that it is optimal to prioritize station 1.*

Final remarks

- Denote prioritizing station 1 by P1 and prioritizing station 2 by P2.
- Benefits of P2:
  - Easy to implement.
  - Follows patient throughout her/his service “cycle”.
- Drawbacks of P2:
  - Restrictive condition.
  - P2 spends highest proportion of time at station 2.

Numerical Study: Prelude

Threshold Policies

- Threshold policy with level T: medical service provider works at station 2 until
  - Station 2 is empty or
  - Number of patients at station 1 reaches T.

Numerical Study: Prelude

Threshold Policies

- Threshold policy with level T: medical service provider works at station 2 until
  - Station 2 is empty or
  - Number of patients at station 1 reaches T.
- Exhaustive Policy (**E**)
- P2 $(T = \infty)$, P1 $(T = 1)$, spend, respectively, highest and least proportion of effort at station 2.
- Between these two extremes are threshold policies with higher thresholds spending more time at station 2.

Optimal Control of Triage and Treatment

- Background
- Modeling Approach
- Numerical Study
- Concluding Remarks

Ongoing and Future Work

Parameter Symbol	Value(s)
μ_1	8.57
μ_2	4.62
β	$0.15, 0.3, 0.5, 0.8$
p	1
R_1	$10, 15$
R_2	20
λ	$0.5, 1.5, 3, 4.5, 6.5, 8.5$

Table: List of Parameters and their values

Parameter Symbol	Value(s)
μ_1	8.57
μ_2	4.62
β	$0.15, 0.3, 0.5, 0.8$
p	1
R_1	10, 15
R_2	20
λ	$0.5, 1.5, 3, 4.5, 6.5, 8.5$

From LMC's TTR.

Table: List of Parameters and their values

Parameter Symbol	Value(s)
μ_1	8.57
μ_2	4.62
β	$0.15, 0.3, 0.5, 0.8$
p	1
R_1	10, 15
R_2	20
λ	$0.5, 1.5, 3, 4.5, 6.5, 8.5$

From LMC's TTR.

Mandelbaum and Zeltyn (2007); Batt and Terwiesch (2013).

Table: List of Parameters and their values

Parameter Symbol	Value(s)
μ_1	8.57
μ_2	4.62
β	$0.15, 0.3, 0.5, 0.8$
p	1
R_1	$10, 15$
R_2	20
λ	$0.5, 1.5, 3, 4.5, 6.5, 8.5$

From LMC's TTR.

Mandelbaum and Zeltyn (2007); Batt and Terwiesch (2013).

$\mu_1 R_1 \leq \mu_2 R_2$; $\mu_1 R_1 > \mu_2 R_2$.

Table: List of Parameters and their values

Parameter Symbol	Value(s)
μ_1	8.57
μ_2	4.62
β	0.15, 0.3, 0.5, 0.8
p	1
R_1	10, 15
R_2	20
λ	0.5, 1.5, 3, 4.5, 6.5, 8.5

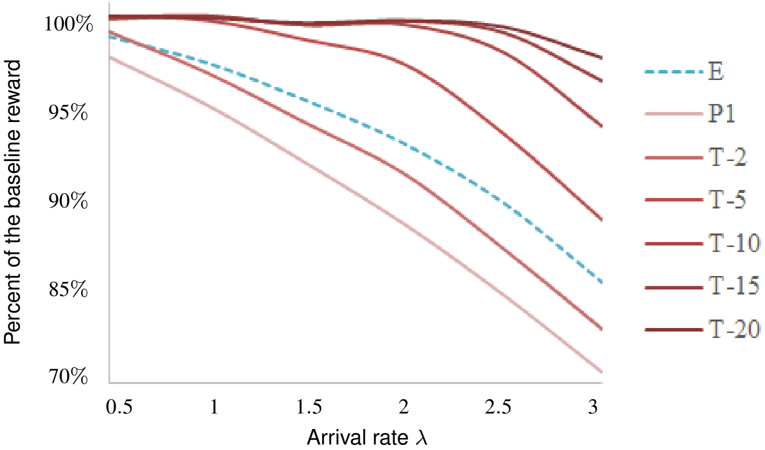
μ_1, μ_2: From LMC's TTR.

β: Mandelbaum and Zeltyn (2007); Batt and Terwiesch (2013).

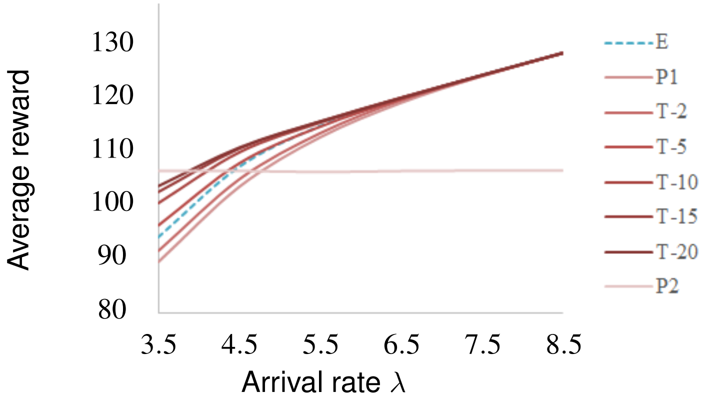
R_1, R_2: $\mu_1 R_1 \leq \mu_2 R_2$; $\mu_1 R_1 > \mu_2 R_2$.

λ = 4.5, 6.5, 8.5: $\frac{1}{\frac{1}{\mu_1} + \frac{1}{\mu_2 + \beta}} \leq \lambda < \mu_1$.

Table: List of Parameters and their values



Percent of the baseline reward ($\beta = 0.8, R_1 = 10$)



Average reward ($\beta = 0.8, R_1 = 15$)

REMARKS

When P2 is stable:

- Decreasing the threshold makes the average reward worse.
  - In **all** instances, P1 $(T = 1)$ performed the worst.
- If $\lambda \in \{0.5, 1\}$, all policies comparable to P2 – within 6% of the optimal
- Similar observations hold for $R_1 = 15$.

Remarks

When P2 is stable:

- Decreasing the threshold makes the average reward worse.
  - In **all** instances, P1 $(T = 1)$ performed the worst.
- If $\lambda \in \{0.5, 1\}$, all policies comparable to P2 – within 6% of the optimal
- Similar observations hold for $R_1 = 15$.

When P2 is not stable, and P1 is used:

- Gains in average reward can be obtained if we are close to stability by using threshold policies but at the cost of larger queue lengths.

RECOMMENDATIONS FOR A TTR SYSTEM

Threshold policies with parameter T - reasonable alternatives to P1 ($T = 1$) and P2 ($T = \infty$)

- P2 is stable
  - If system is lightly loaded, no significant loss of optimality.
  - If system is highly loaded, there is significant loss of optimality.

RECOMMENDATIONS FOR A TTR SYSTEM

Threshold policies with parameter T - reasonable alternatives to P1 $(T = 1)$ and P2 $(T = \infty)$

- P2 is stable
  - If system is lightly loaded, no significant loss of optimality.
  - If system is highly loaded, there is significant loss of optimality.
- P2 is unstable – impractical
  - Average reward of alternative policies are not too different – a provider might consider policies with the lowest average total number in the system, say.

ADDITIONAL CHALLENGES FROM THE ER

- Arrival processes are non-stationary (time-dependent) and often periodic
  - Replace homogeneous Poisson process with a non-homogeneous Poisson process or Markov modulated process
- Patients/customers are impatient
  - Models should include abandonments at **both stages**
- Health can be deteriorating
  - Service times are usually not exponential.

Additional Challenges from the ER

- Arrival processes are non-stationary (time-dependent) and often periodic
  - Replace homogeneous Poisson process with a non-homogeneous Poisson process or Markov modulated process
- Patients/customers are impatient
  - Models should include abandonments at **both stages**
- Health can be deteriorating
  - Service times are usually not exponential.

The Road Ahead

To address these and other challenges relevant to healthcare.

AN ADMISSION CONTROL PROBLEM

Background: Patients having different severity levels require medical care at the E.R.

Question: How to control admissions into an E.R. with limited resources (e.g. beds, examination rooms, or medical equipment)?

An Admission Control Problem

Background: Patients having different severity levels require medical care at the E.R.

Question: How to control admissions into an E.R. with limited resources (e.g. beds, examination rooms, or medical equipment)?

Modeling Approach:

- Can be modeled as an admission control problem using CTMDP.

Challenges and Considerations:

- Challenges highlighted in the previous slide.

Ambulance Diversion Policies

Design and Analysis

Background: Hospital overcrowding leads managers to request that incoming ambulances be sent to neighboring hospitals, a phenomenon known as *ambulance diversion*.

Questions: When should a hospital go on ambulance diversion? How should this be affected by conditions at the other hospitals in the region?

Ambulance Diversion Policies

Design and Analysis

Background: Hospital overcrowding leads managers to request that incoming ambulances be sent to neighboring hospitals, a phenomenon known as *ambulance diversion*.

Questions: When should a hospital go on ambulance diversion? How should this be affected by conditions at the other hospitals in the region?

Modeling Approach:

- Can be modeled as a routing control problem using CTMDP.

Challenges and Considerations:

- Set-up and transportation times.
- Curse of dimensionality — May require approximate dynamic programming and simulation.

Thank you!