Optimal Scheduling of Proactive Care with Patient Deterioration

Yue Hu (Columbia Business School)

Joint work with
Carri Chan (Columbia Business School)
and
Jing Dong (Columbia Business School)

Motivation

- Healthcare is a limited resource environment where scarce capacity is often reserved for the most severe patients
- With the advancement of data availability and analytical methods, we can develop more accurate predictive models

$\Longrightarrow$ Proactive Care / Preventative Care

Motivation

Early Warning Systems in hospital:

- Prognostic systems that monitor hospitalized patients and prompt alarms for intensive care unit admissions
- Early warning systems for cardiovascular risk
- Predictive models for hospital-acquired infection
- Readmission risk at discharge

"Recent systematic reviews have demonstrated that Early Warning System based alarms only marginally improve outcomes while substantially increasing physician and nursing workloads"

Key Consideration

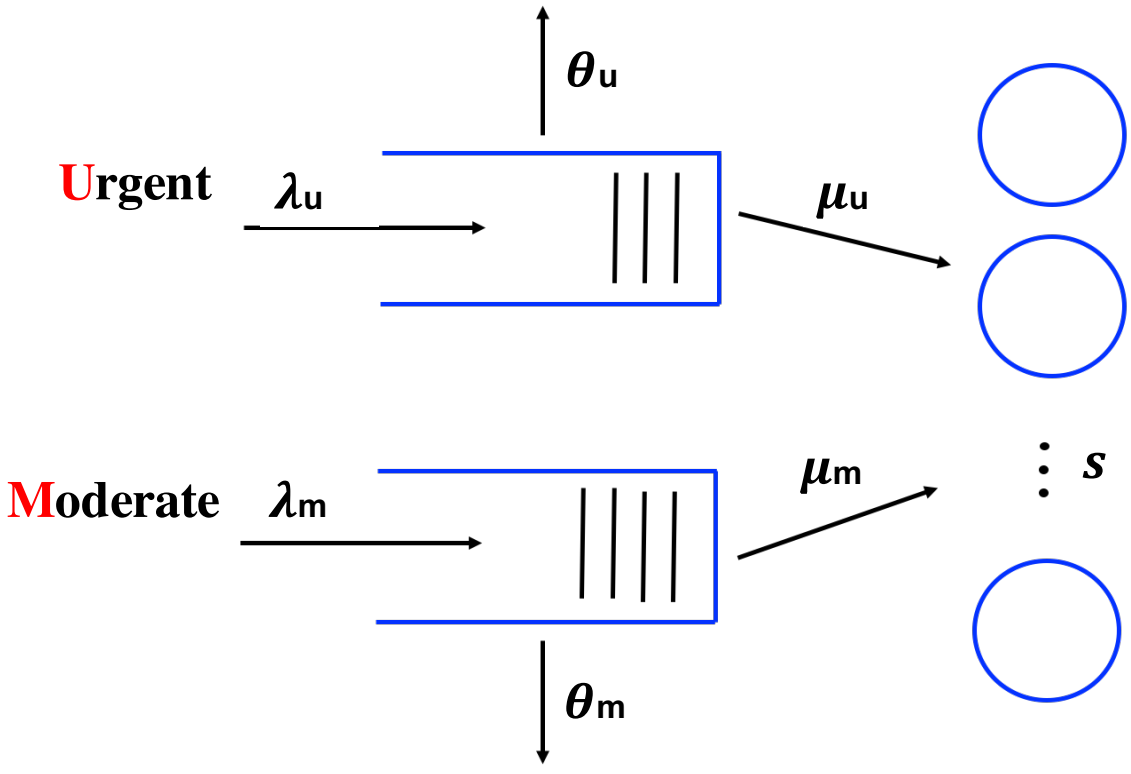
When proactive care is an option, we face the tension!

- Providing care for patients when they are less critical could mean that fewer resources are necessary to return them to a healthy and stable state
- With limited resources, providing proactive care may delay treatment for the more critical patients
- Some of the less critical patients may become stable without ever needing critical care

Goal: Develop a better understanding of these tradeoffs and derive an optimal scheduling policy for proactive care

The Model

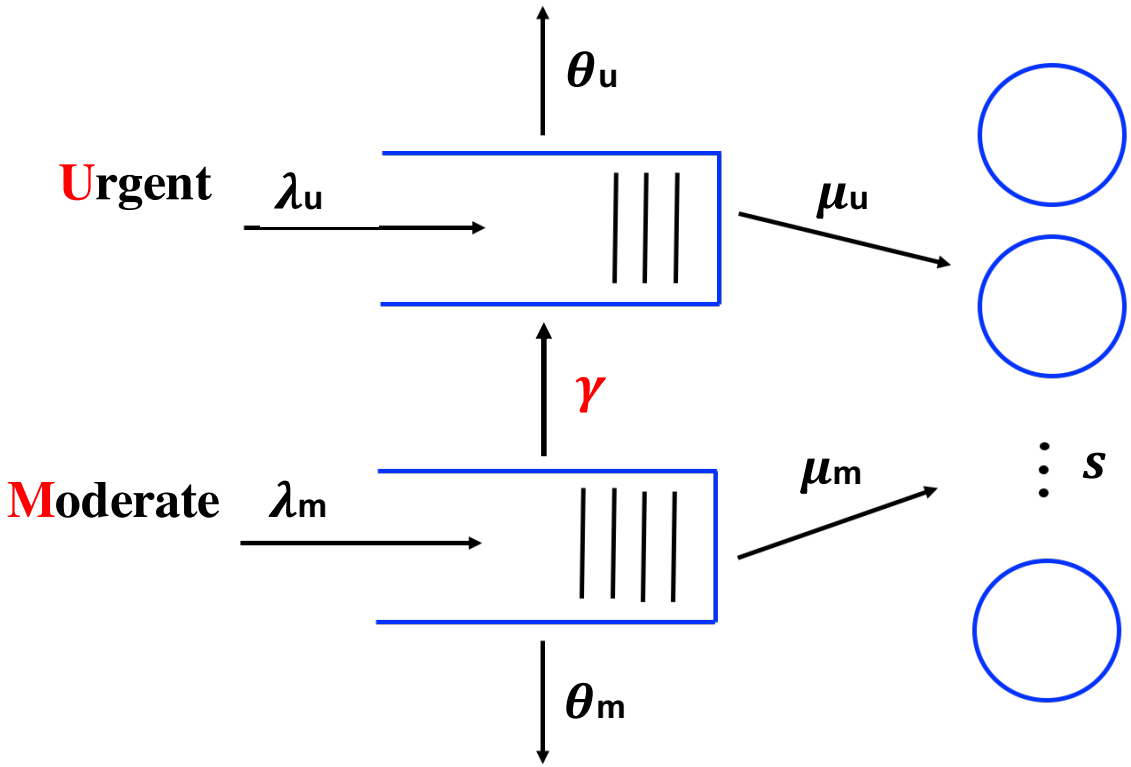
A stochastic queueing network where two queues are served by s servers



- Stationary arrival process of jobs with rate λ_i to queue $i \in \{u, m\}$
- IID service times with rate μ_i at queue $i \in \{u, m\}$, $\mu_u < \mu_m$

The Model

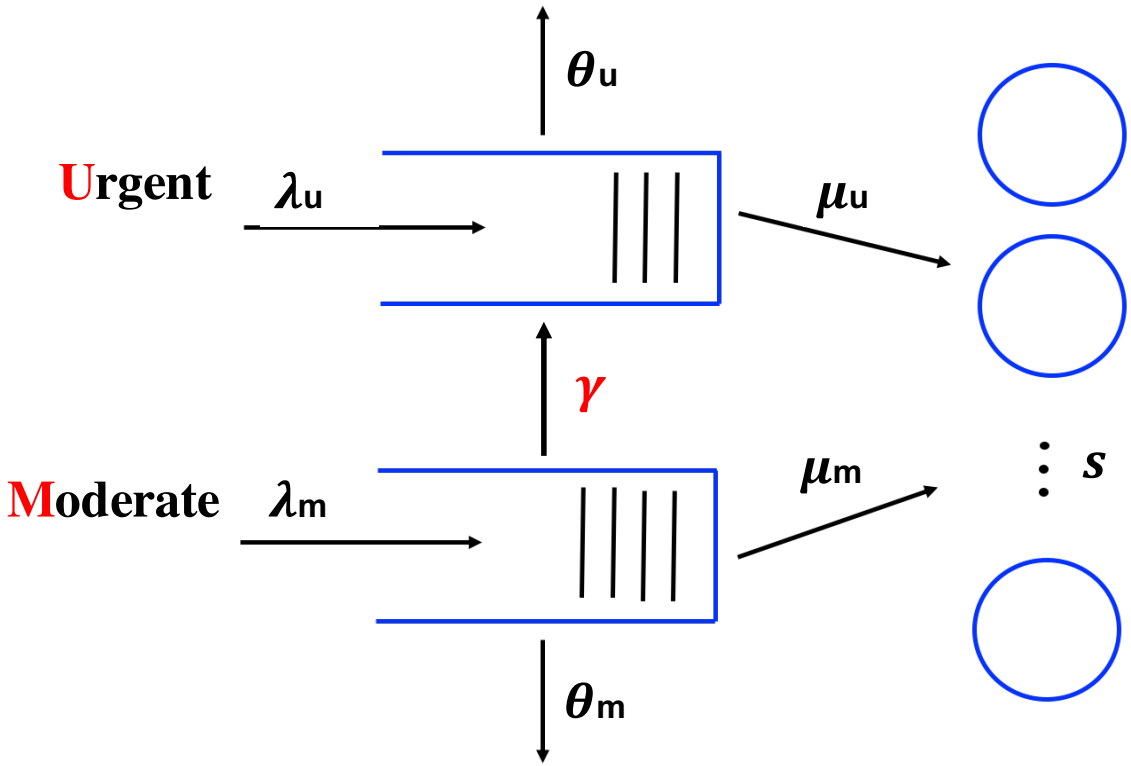
A stochastic queueing network where two queues are served by s servers



- Stationary arrival process of jobs with rate λ_i to queue $i \in \{u, m\}$
- IID service times with rate μ_i at queue $i \in \{u, m\}$, $\mu_u < \mu_m$
- Delayed moderate patients become urgent at rate γ according to a stationary arrival process

The Model

A stochastic queueing network where two queues are served by s servers



$\frac{\gamma}{\gamma+\theta_m}$:

- proportion of moderate patients who deteriorate into urgent ones
- true positive rate of the early warning system
- prediction accuracy

Research Questions

- When the system is in the "normal" state of operation, what is the optimal scheduling rule?

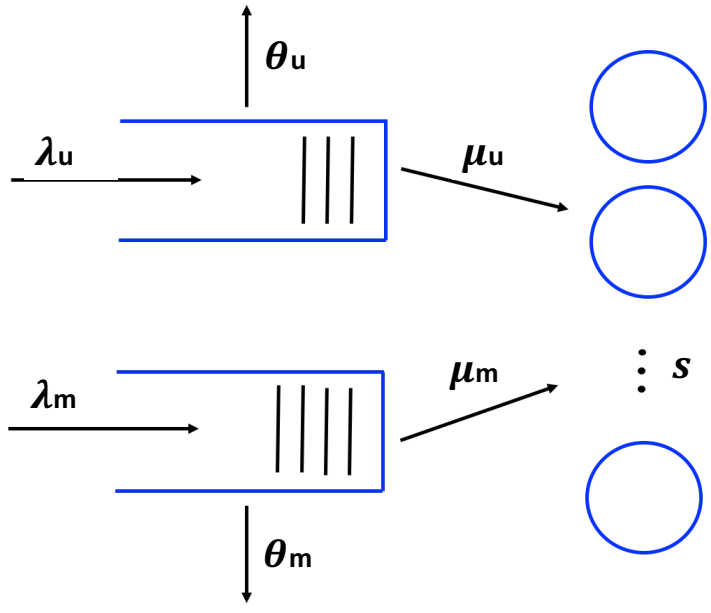
  $\Longrightarrow$ Long-run average performance / equilibrium performance

- When random shocks (disease outbreak or mass casualty events) bring the system far from its normal state of operation, what is the optimal scheduling policy to bring the system back to normal?

  $\Longrightarrow$ Transient performance

Challenges

- Overloaded regime: the $c\mu/\theta$ rule is optimal (Atar et al. (2011))
- Limiting heavy-traffic regime:
  - the optimal control is the solution to the associated Hamilton-Jacobi-Bellman equation (Harrison and Zeevi (2003))
- Special Case: transient two-queue fluid system (Larrañaga (2015))



Related Literature

Patient Deterioration and Customer Slowdown:

Sheridan et al. (1999), Richardson (2002), Liew et al. (2003), Siegmeth et al. (2005), Chalfin et al. (2007), Chan et al. (2008), Renaud et al. (2009)

Proactive Service and Multi-class Queues with Dynamic Class Type:

Ozekici and Pliska (1991), Ormeci et al. (2015), Sun et al. (2017), Hu et al. (2018), Xu and Chan (2016), Delana et al. (2019) Akan et al. (2012), Xie et al. (2017), Cao and Xie (2016), Down and Lewis (2010)

Transient Queueing Control:

Abate and Whitt (1988, 2006), Hartl et al. (1995), Honnappa et al. (2015), Larranaga et al. (2013), Larranaga (2015)

Roadmap

- Fluid approximation
- Optimal scheduling policy to minimize the long-run average holding cost
- Optimal scheduling policy to minimize the transient holding cost

A Fluid Approximation to Simplify the Problem

Consider a piecewise affine dynamical system characterized by

$$dq_u(t) = \lambda_u - \mu_u z_u(t) - \theta_u q_u(t) + \gamma q_m(t)$$
$$dq_m(t) = \lambda_m - \mu_m z_m(t) - \theta_m q_m(t) - \gamma q_m(t)$$

where $z_i(t)$ is the amount of capacity devoted to serving Class i patients satisfying

$$z_i(t) \geq 0, \quad i = u, m, \ t \geq 0$$
$$z_u(t) + z_m(t) \leq s, \quad t \geq 0$$
$$dq_i(t) \geq 0 \text{ whenever } q_i(t) = 0, \quad i = u, m, \ t \geq 0$$

The set of admissible controls, denoted by $\mathcal{F}$, are Markov, non-anticipatory and preemptive.

Long-run Average Optimization - Definition

For any staffing level s, the long-run optimization problem is

Problem (Long-run average optimization)

$$\min_{\pi \in \mathcal{F}} \limsup_{T \to \infty} \frac{1}{T} \int_0^T \left(c_u q_u(t) + c_m q_m(t) \right) dt$$

In a Markovian setting, the cost rates c_u and c_m incorporates unit-time holding cost, fixed abandonment cost, and fixed degradation cost.

Theorem (Optimal long-run scheduling policy)

The modified $c\mu/\theta$-rule (a simple static index policy) is optimal for the long-run average optimization problem

Long-run Average Optimization

- Let P_u and P_m denote the strict priority rule to urgent and moderate patients
- The modified $c\mu/\theta$-rule decides when to use P_u or P_m in the descending order of the modified $c\mu/\theta$-index
- The modified $c\mu/\theta$-index for urgent patients is

$$\frac{c_u \mu_u}{\theta_u}$$

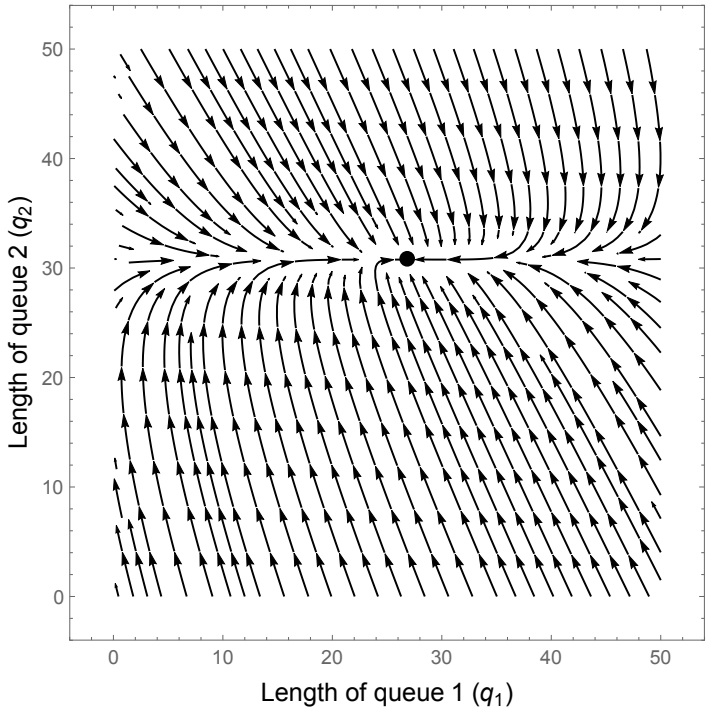
- The modified $c\mu/\theta$-index for moderate patients is

$$\frac{c_m}{\gamma + \theta_m}\mu_m + \frac{\frac{\gamma}{\gamma+\theta_m}c_u}{\theta_u}\mu_m$$

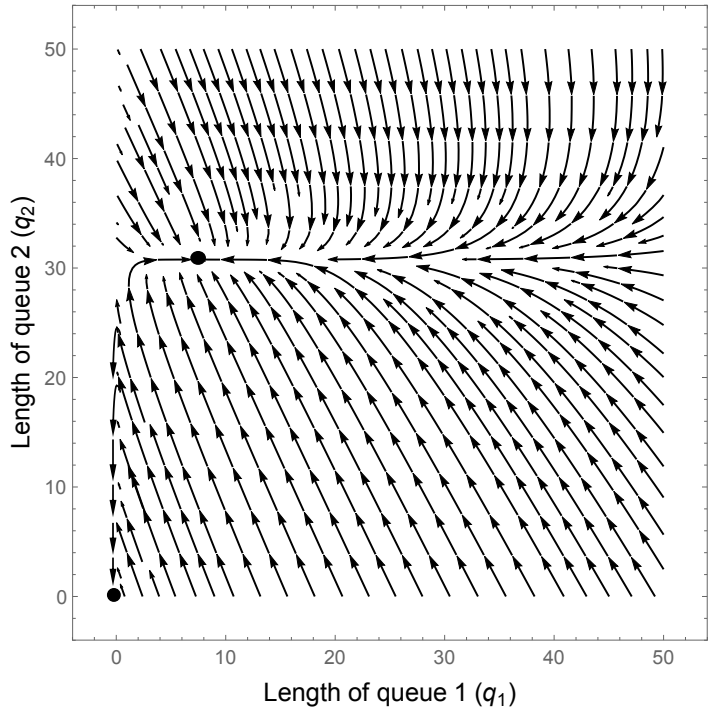
Long-run Average Optimization

We derive the optimal long-run scheduling policy by analyzing the long-run behavior of the fluid process under the strict priority rules

Example: fluid queue length process under P_u



(a) Case 1. $\mu_u > \frac{\gamma}{\gamma+\theta_m}\mu_m$: Globally asymptotically stable equilibrium

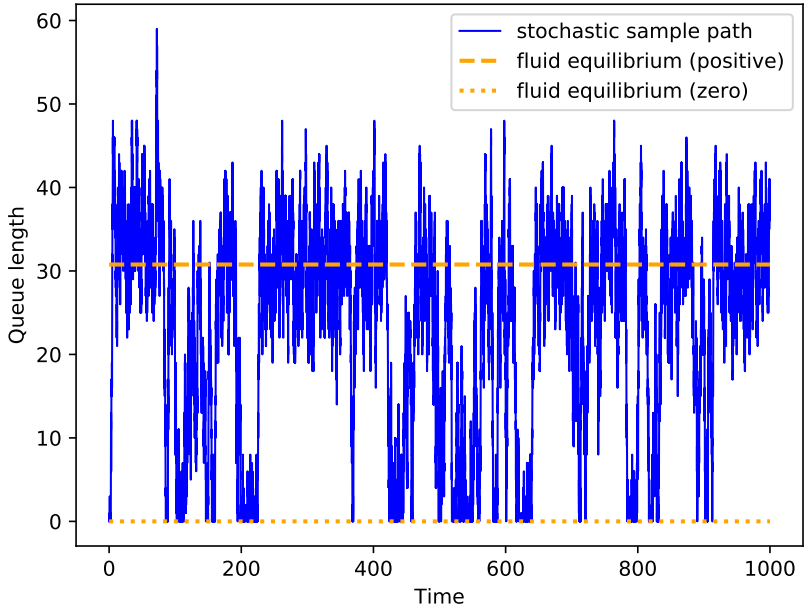


(b) Case 2. $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$: Locally asymptotically stable equilibrium

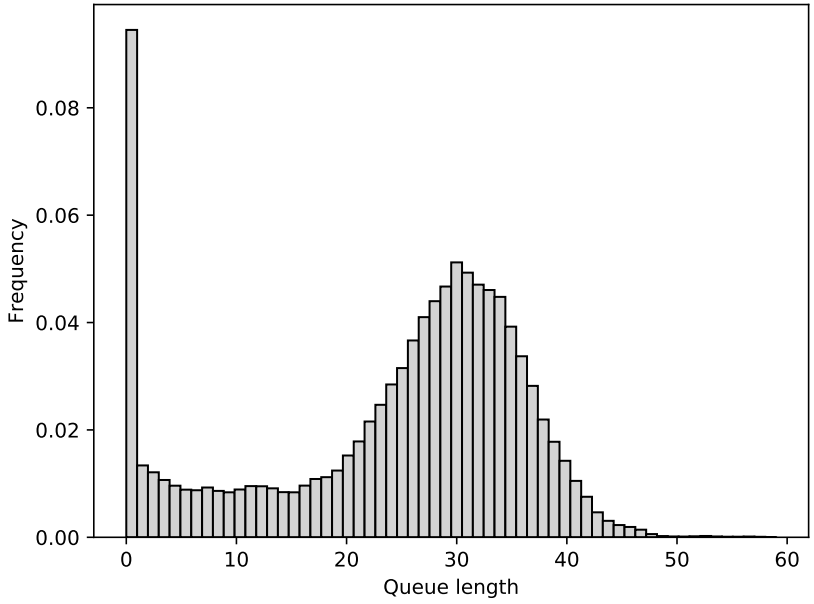
Long-run Average Optimization

For the stochastic system, bi-stability means that the queue length process fluctuates between the two fluid equilibria infinitely often

Example: stochastic queue length process of moderate patients under P_u



(c) Sample path

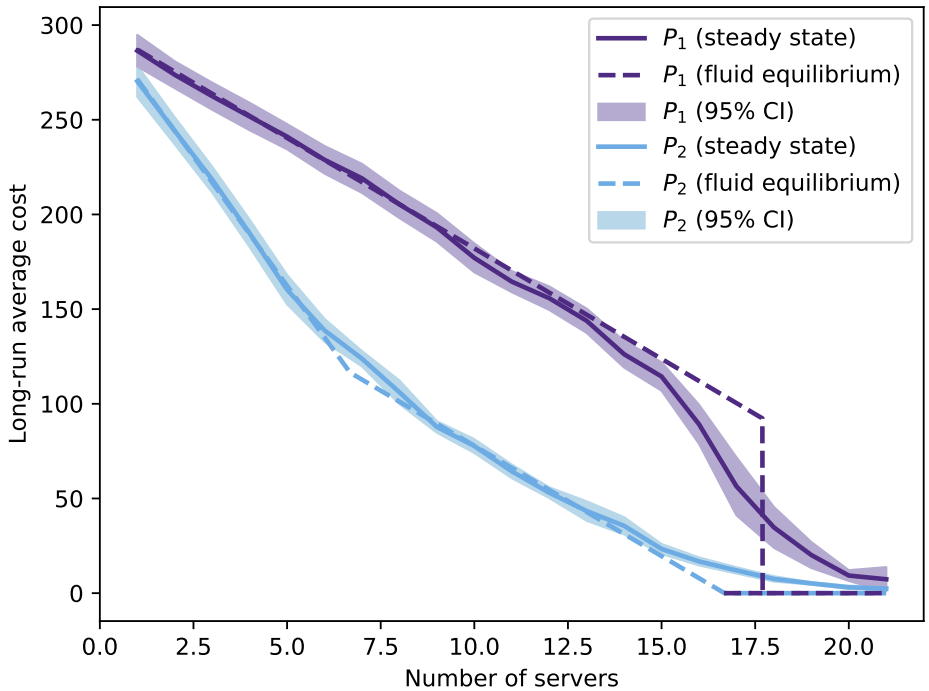


(d) Histogram

The Modified $c\mu/\theta$-Rule

Optimal long-run scheduling policy in Case 2: $\mu_u < \frac{\gamma}{\gamma+\theta_m}\mu_m$

- In this case, it always holds that $\frac{c_u}{\theta_u}\mu_u > \frac{c_m}{\theta_m+\gamma}\mu_m + \frac{\frac{\gamma}{(\theta_m+\gamma)}c_u}{\theta_u}\mu_m$



Transient Optimal Control - Definition

- Assume $s > \lambda_u/\mu_u + +\lambda_m/\mu_m$, so that there is sufficient service capacity to empty the system in a finite amount of time given any initial condition
- Define the first system empty time $\tau := \inf\{t \geq 0 : q_u(t) + q_m(t) = 0\}$

Problem (Transient optimal control)

$$\begin{aligned}
\min_{\pi \in \mathcal{F}} \quad & \int_0^\tau (c_u q_u(t) + c_m q_m(t))\, dt \\
s.t. \quad & dq_u(t) = \lambda_u - \mu_u z_u(t) - \theta_u q_u(t) + \gamma q_m(t) \\
& dq_m(t) = \lambda_m - \mu_m z_m(t) - (\gamma + \theta_m) q_m(t) \\
& z_u(t) + z_m(t) \leq s \\
& z_u(t), z_m(t), q_u(t), q_m(t) \geq 0
\end{aligned}$$

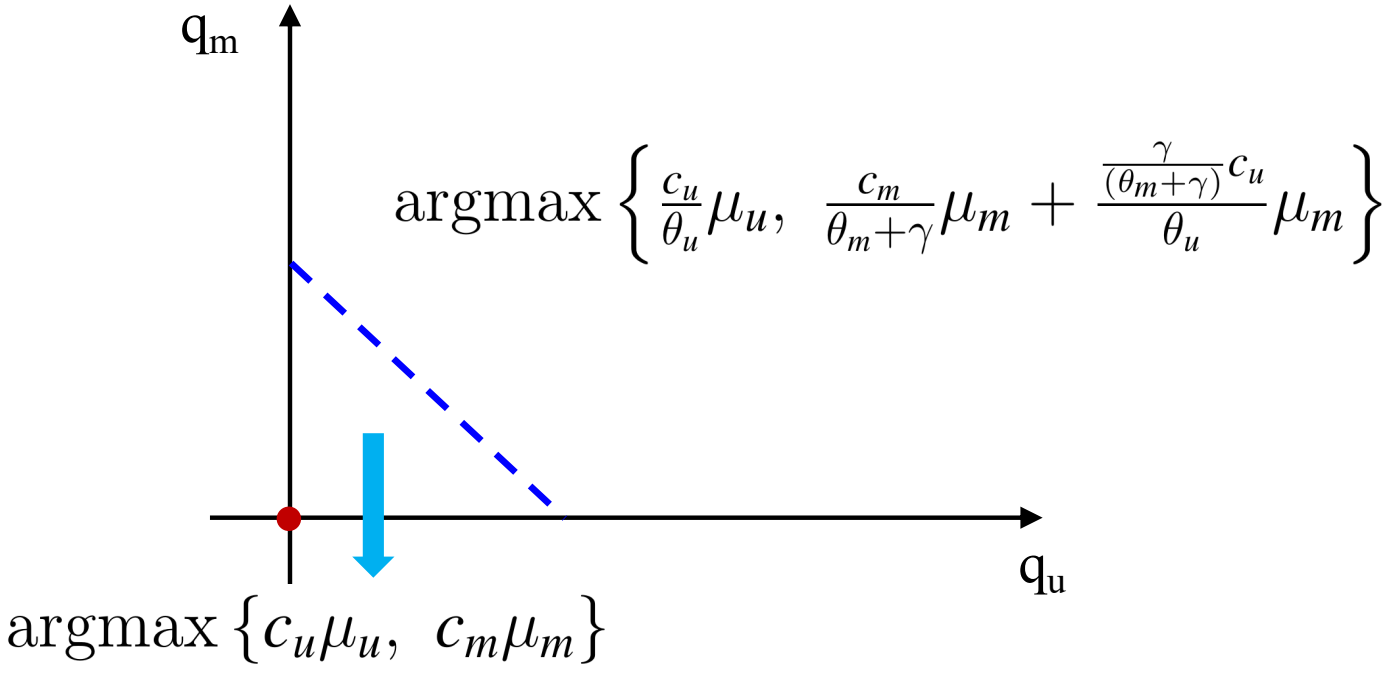
Transient Optimal Control

Theorem (Optimal transient scheduling policy)

For the transient optimal control problem,

- *the modified $c\mu/\theta$-rule is optimal when the states are far from the origin*
- *the $c\mu$-rule is optimal when the states are close to the origin*

Furthermore, the optimal control switches priority at most once



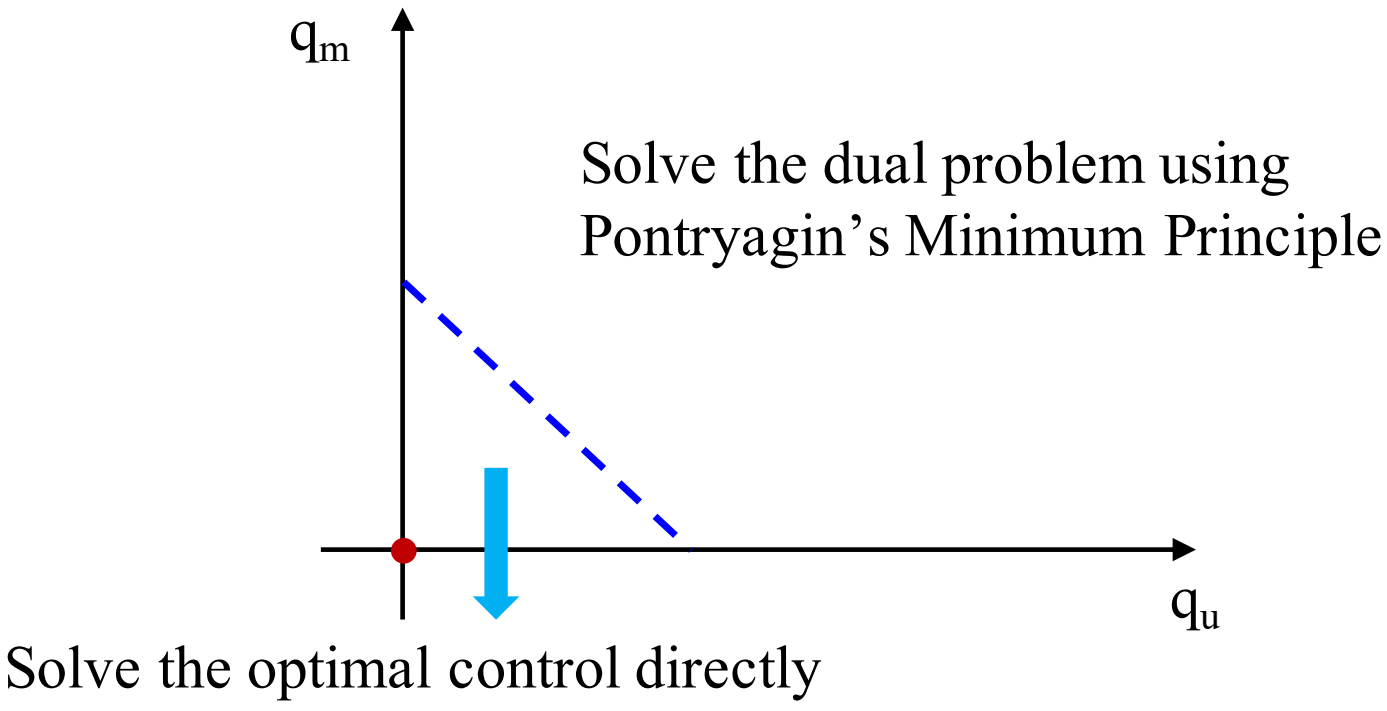
Transient Optimal Control

Theorem (Optimal transient scheduling policy)

For the transient optimal control problem,

- *the modified $c\mu/\theta$-rule is optimal when the states are far from the origin*
- *the $c\mu$-rule is optimal when the states are close to the origin*

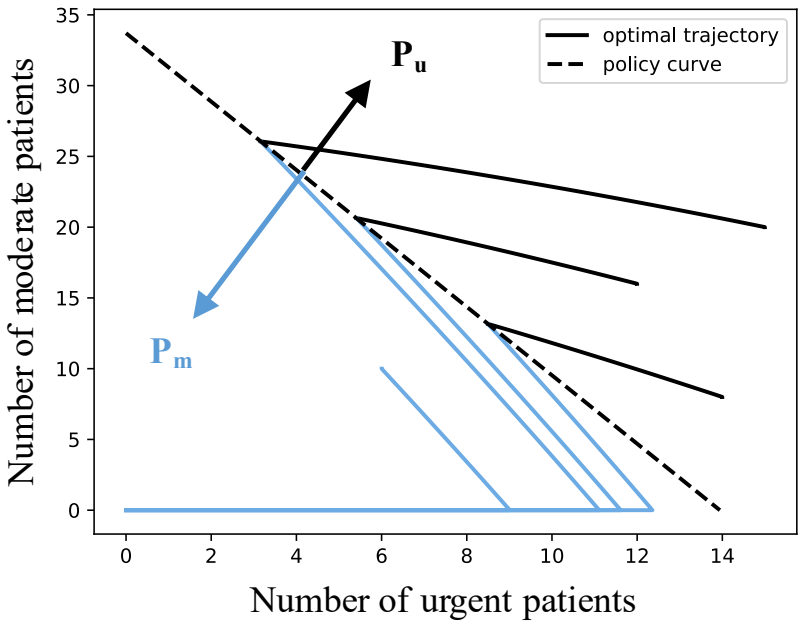
Furthermore, the optimal control switches priority at most once



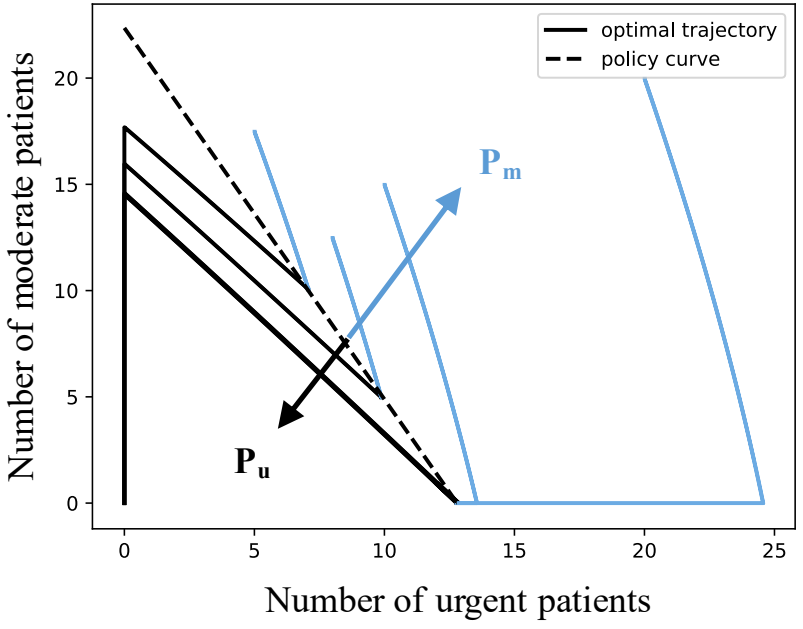
Transient Optimal Control

Example: Optimal transient state trajectories

The optimal control switches from the modified $c\mu/\theta$-rule to the $c\mu$-rule when the state trajectory crosses the policy curve $\mathcal{P}$



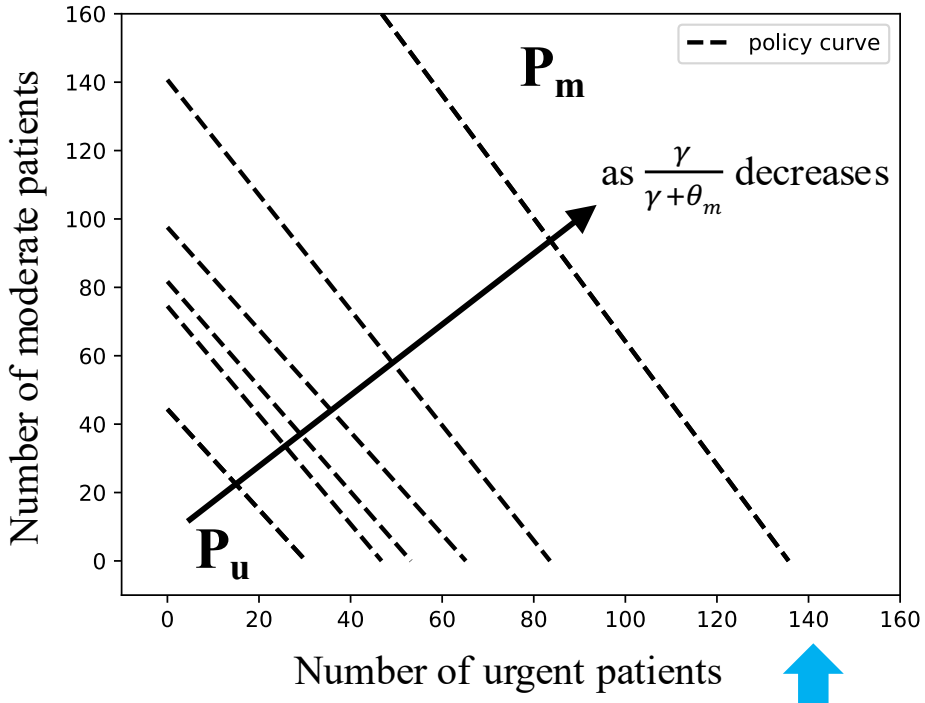
(a) $c\mu$-rule: P_m, modified $c\mu/\theta$-rule: P_u



(b) $c\mu$-rule: P_u, modified $c\mu/\theta$-rule: P_m

Transient Optimal Control

Sensitivity analysis: The policy curve $\mathcal{P}$ for switching from P_u to P_m



Prediction accuracy $\frac{\gamma}{\gamma+\theta_m}$

Summary

- We propose a two-class multi-server queueing model to study the potential of proactive care with degrading class types
- We consider a fluid approximation and characterize optimal long-run and transient scheduling policies
  - When the system is in the "normal state of operation, what is the optimal priority rule?
    Ans: The modified $c\mu/\theta$-rule
  - When random shocks (disease outbreak or mass casualty events) bring the system far from its normal state of operation, what is the optimal scheduling policy to bring the system back to normal?
    Ans: The modified $c\mu/\theta$-rule (far), the $c\mu$-rule (close)

Thank You