Aalto University
School of Science
and Technology

Control and optimization of single-server queues: the Gittins index approach revisited

Samuli Aalto, Aalto University, Finland

in cooperation with

Urtzi Ayesta, BCAM, Spain

Rhonda Righter, UC Berkeley, USA

NET-COOP 2010
29 Nov – 1 Dec 2010
Ghent, Belgium

Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Problem formulation

- Transient system
  - Given a single-server queue
    with n IID jobs and service time distribution $F(x)$,
    what is the optimal non-anticipating service policy
    so that the mean delay is minimized?
- Dynamic system
  - Given an M/G/1 queue
    with arrival rate λ and service time distribution $F(x)$,
    what is the optimal non-anticipating service policy
    so that the mean delay is minimized?

Optimality of SRPT

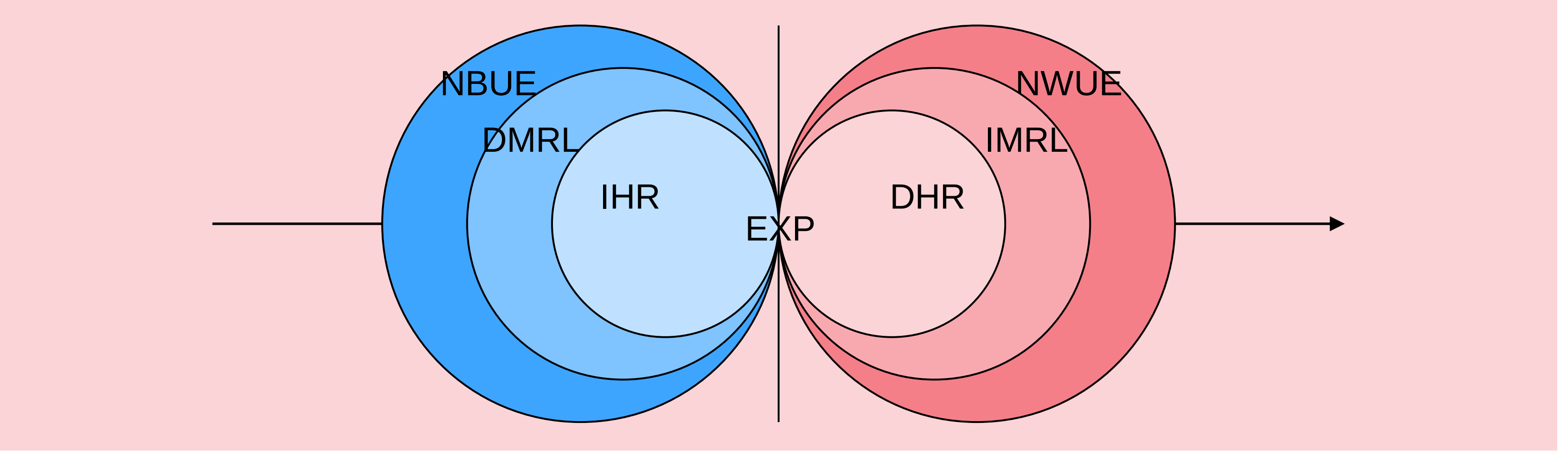
- For both problems,
  the optimal anticipating policy is SRPT,
  but it requires exact information about the service times

- However,
  we are looking for the optimal non-anticipating policy,
  since we are not given the service times, only their distribution is known

- Note also that we allow preemptions

Service time distribution classes

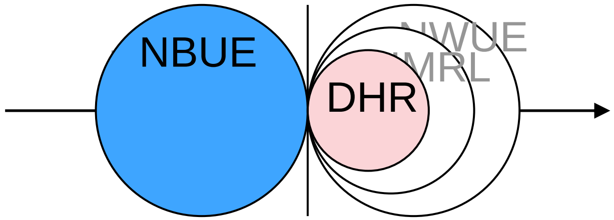
- Service times are
  - IHR [DHR] if hazard rate $h(x)$ is increasing [decreasing]
  - DMRL [IMRL] if $MRL(x)$ is decreasing [increasing]
  - NBUE [NWUE] if $MRL(0) \geq [\leq]\ MRL(x)$



Known optimality results for non-anticipating policies

- Any MAS (a.k.a. NPR) is optimal for NBUE service times
  Righter, Shanthikumar and Yamazaki (1990)

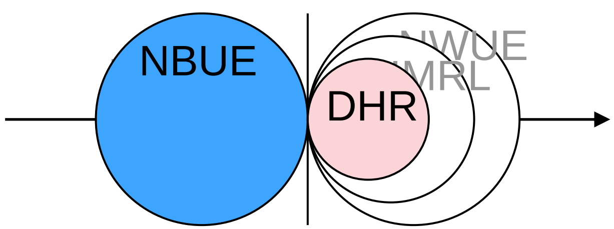
  – MAS = Most Attained Service
  – NPR = Non-Pre-emptive
  – FIFO = First In First Out
  – SIRO = Service In Random Order



Aalto University
School of Science
and Technology

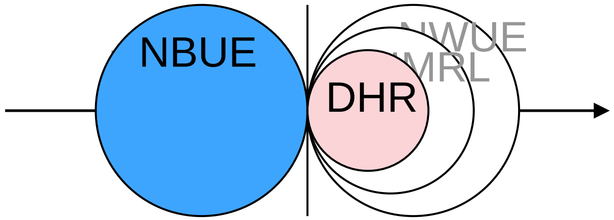
Known optimality results for non-anticipating policies

- Any MAS (a.k.a. NPR) is optimal for NBUE service times
  Righter, Shanthikumar and Yamazaki (1990)
- LAS (a.k.a. FB) is optimal for DHR service times
  Righter and Shanthikumar (1989)
  - LAS = Least Attained Service
  - FB = Foreground Background



Known optimality results for non-anticipating policies

- Any MAS (a.k.a. NPR) is optimal for NBUE service times
  Righter, Shanthikumar and Yamazaki (1990)
- LAS (a.k.a. FB) is optimal for DHR service times
  Righter and Shanthikumar (1989)
- Any GI is optimal for any service time distribution
  Sevcik (1974), Klimov (1974,1978), Gittins (1989), Yashkov (1992)



Gittins index policy

- Definition:
  - Gittins index policy (GI) gives service to the job i with the highest Gittins index $G(a_i)$.

- Observations:
  - GI is not necessary unique
  - Any MAS is a GI if and only if $G(a) \geq G(0)$ for all a.
  - LAS is a GI if and only if $G(a)$ is decreasing for all a.

Example

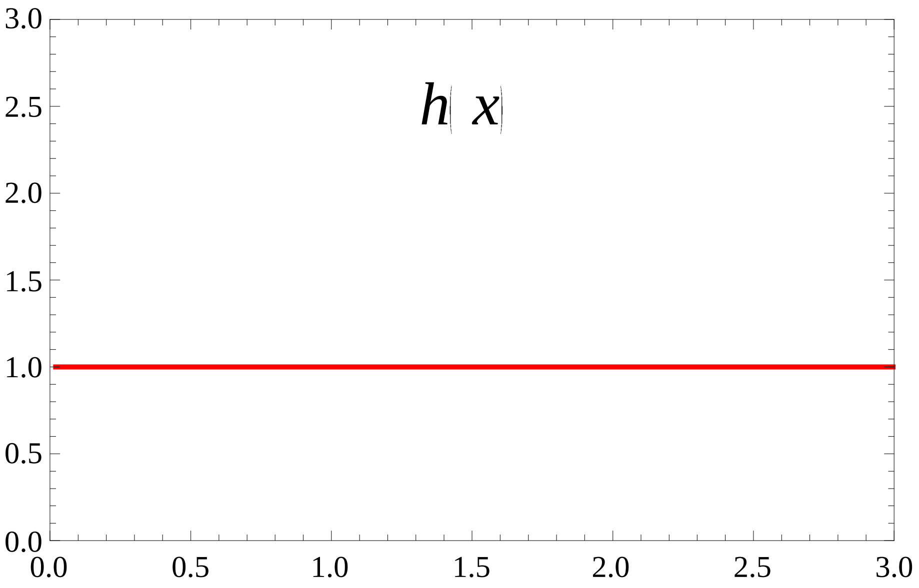
$$n = 3$$

Hazard rate h(x)

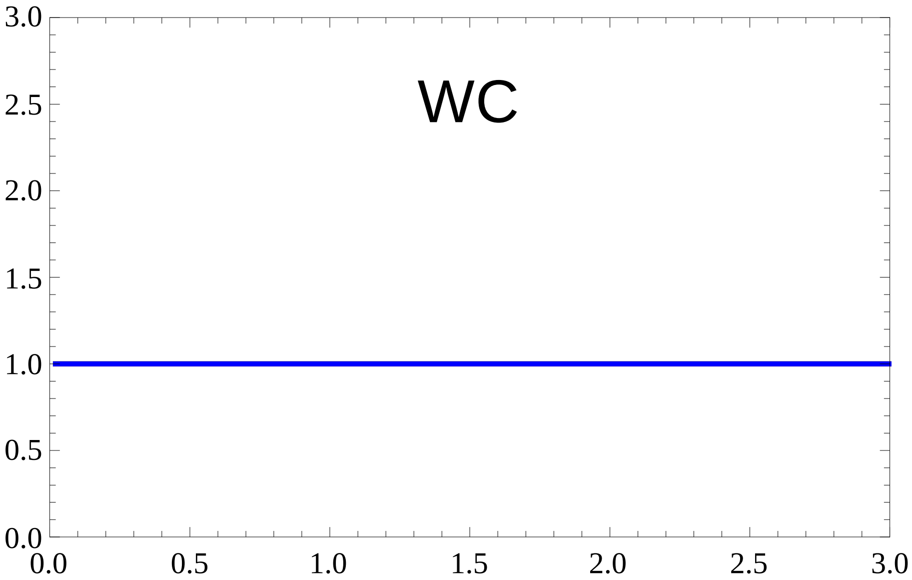
$$F(x) = \int_0^x f(y)dy, \qquad h(x) = \frac{f(x)}{1 - F(x)}$$

Example 1
Constant hazard rate

$$h(x) = 1$$

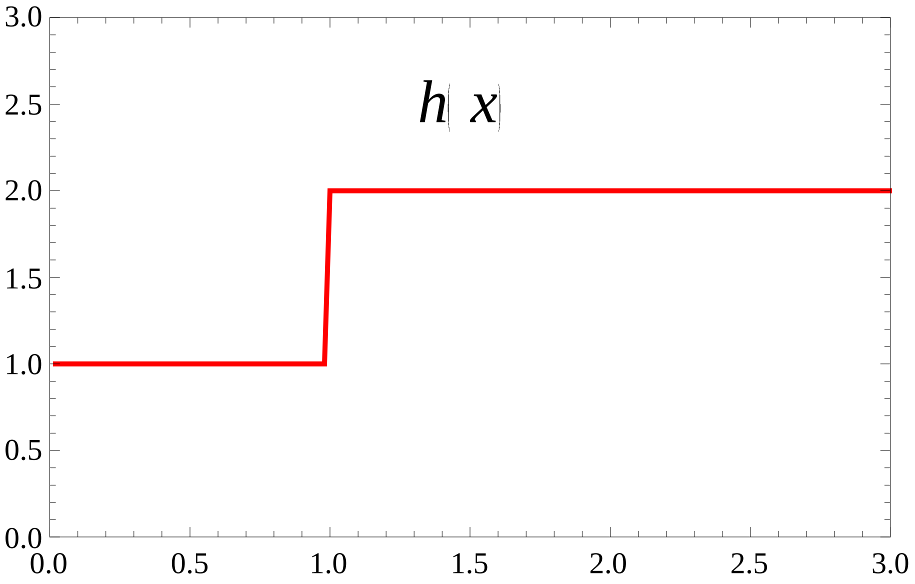


Example 1
Constant hazard rate

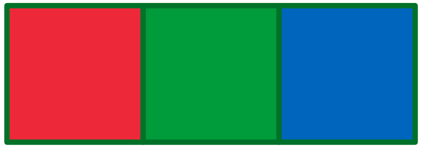


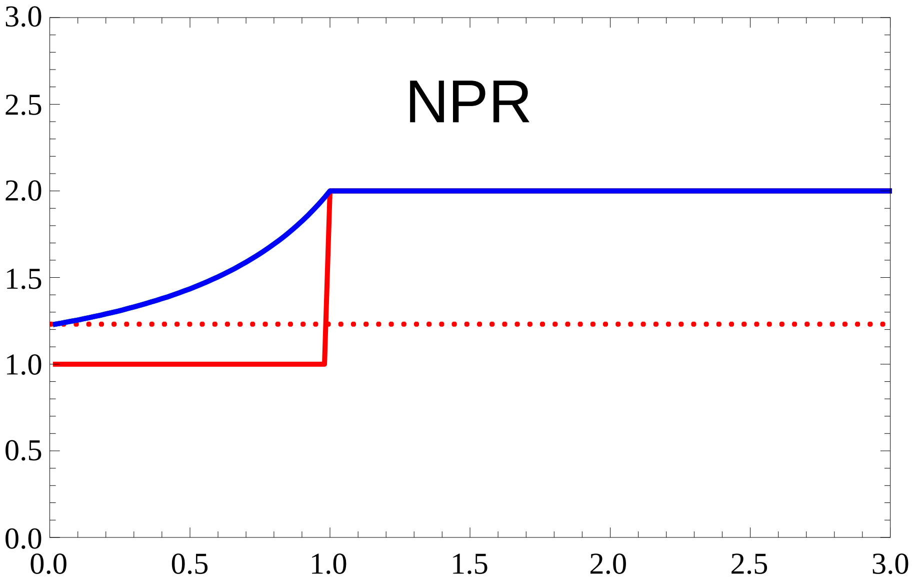
Example 2
Increasing hazard rate

$$h(x) = \begin{cases} 1, & x < 1 \\ 2, & x \geq 1 \end{cases}$$



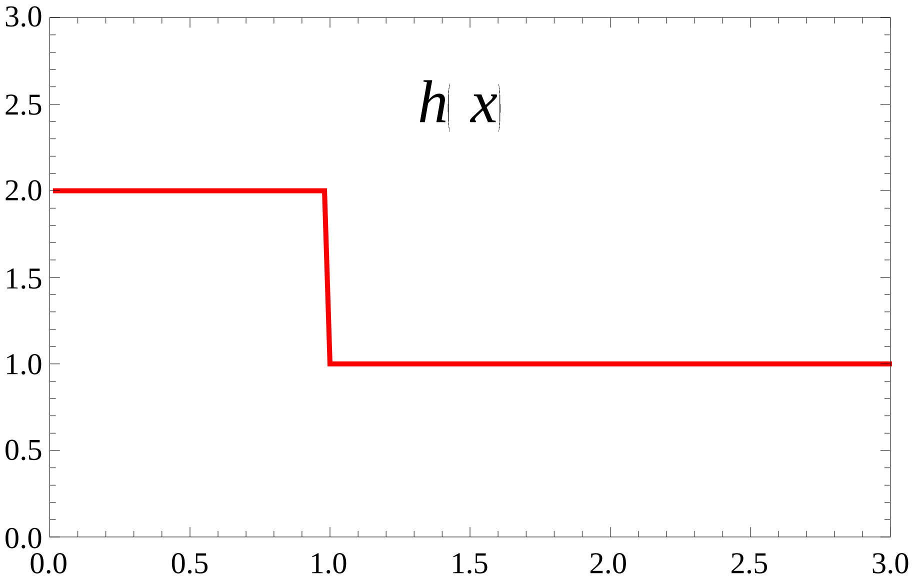
Example 2
Increasing hazard rate



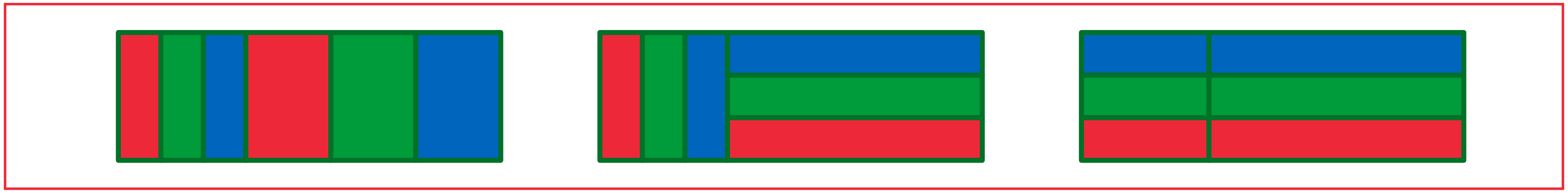


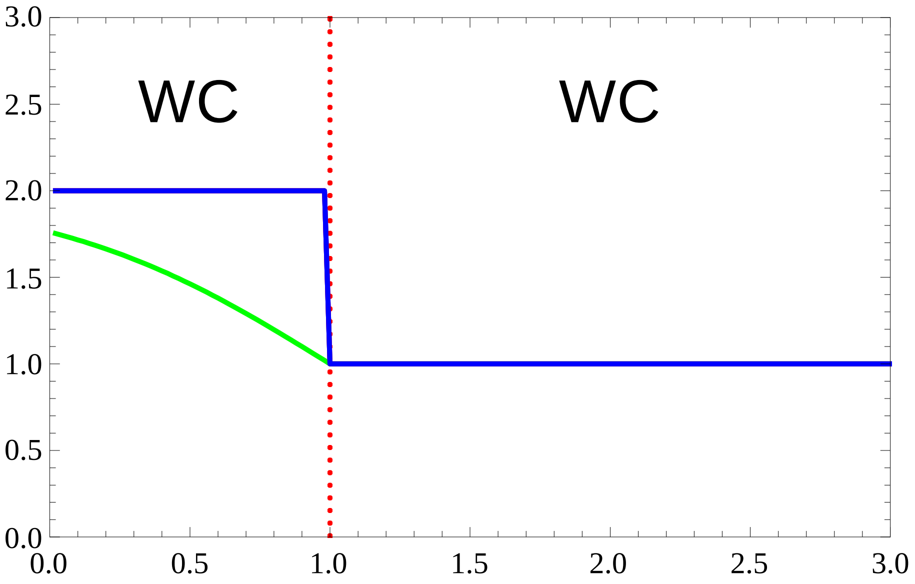
Example 3
Decreasing hazard rate

$$h(x) = \begin{cases} 2, & x < 1 \\ 1, & x \geq 1 \end{cases}$$



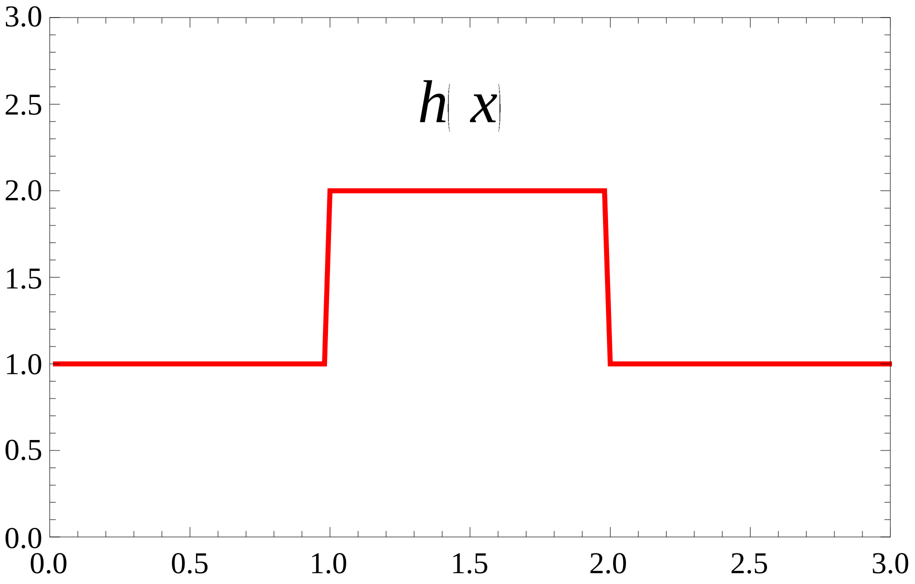
Example 3
Decreasing hazard rate





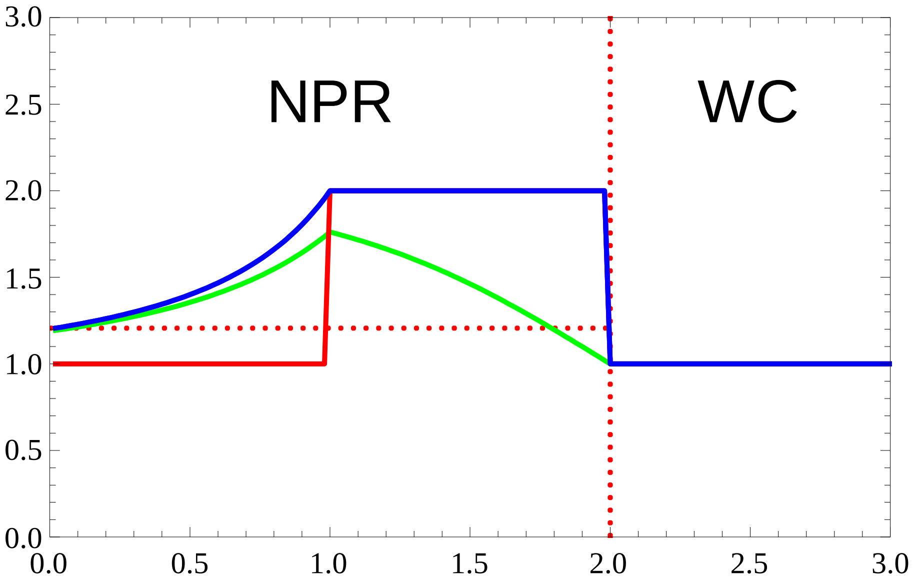
Example 4
Increasing-decreasing hazard rate

$$h(x) = \begin{cases} 1, & x < 1,\ x > 2 \\ 2, & 1 \le x < 2 \end{cases}$$



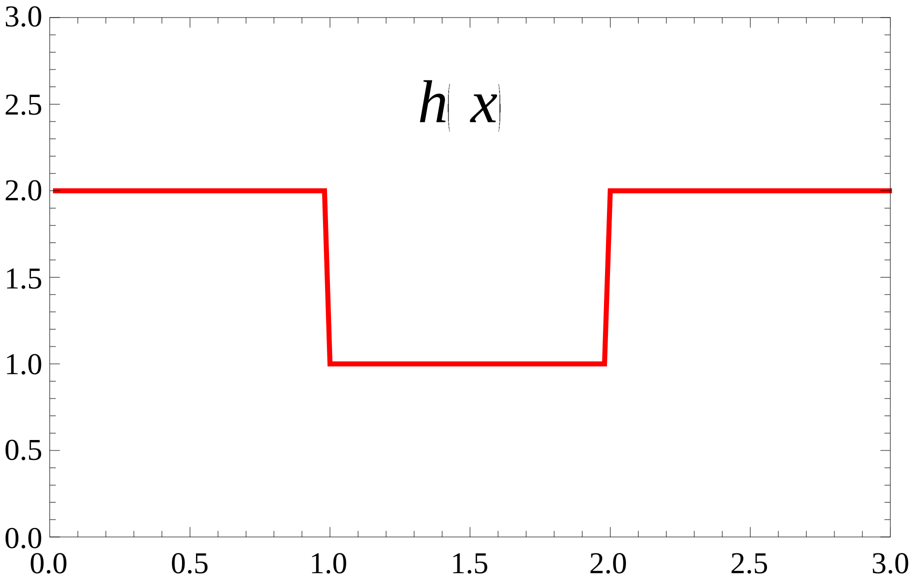
Aalto University
School of Science
and Technology

Example 4
Increasing-decreasing hazard rate

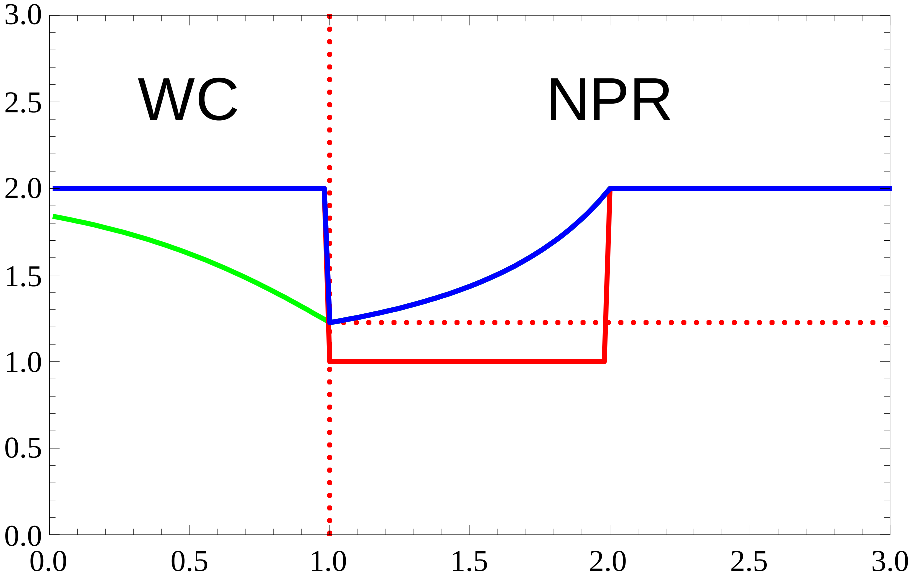


Example 5
Decreasing-increasing hazard rate

$$h(x) = \begin{cases} 2, & x < 1,\ x > 2 \\ 1, & 1 \le x < 2 \end{cases}$$



Example 5
Decreasing-increasing hazard rate



Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Hazard rate

- Service time distribution:

$$F(x) = P\{S \le x\}, \quad \overline{F}(x) = 1 - F(x) > 0 \ \text{ for all } x$$

- Density function:

$$f(x) = P\{S \in dx\} \ \text{ continuous}$$

- Hazard rate function:

$$h(x) \triangleq \lim_{\Delta \to 0} \frac{1}{\Delta} P\{S - x \le \Delta \mid S > x\} = \frac{f(x)}{\overline{F}(x)}$$

Aalto University
School of Science
and Technology

Inverse MRL

- Remaining service time distribution:

$$P\{S - x \le y \mid S > x\} = \frac{\bar{F}(x) - \bar{F}(x+y)}{\bar{F}(x)}$$

- Mean residual lifetime (MRL) function:

$$E[S - x \mid S > x] = \frac{\int_x^{\infty} \bar{F}(y)dy}{\bar{F}(x)}$$

- Inverse MRL function:

$$H(x) \triangleq \frac{1}{E[S - x \mid S > x]} = \frac{\bar{F}(x)}{\int_x^{\infty} \bar{F}(y)dy} = \frac{\int_x^{\infty} f(y)dy}{\int_x^{\infty} \bar{F}(y)dy}$$

Gittins index

- Gittins index for a job with age a:

$$G(a) \triangleq \sup_{\Delta \geq 0} J(a,\Delta)$$

- Optimal service quota for a job with age a :

$$\Delta^*(a) \triangleq \sup\{\Delta \geq 0 \mid J(a,\Delta) = G(a)\}$$

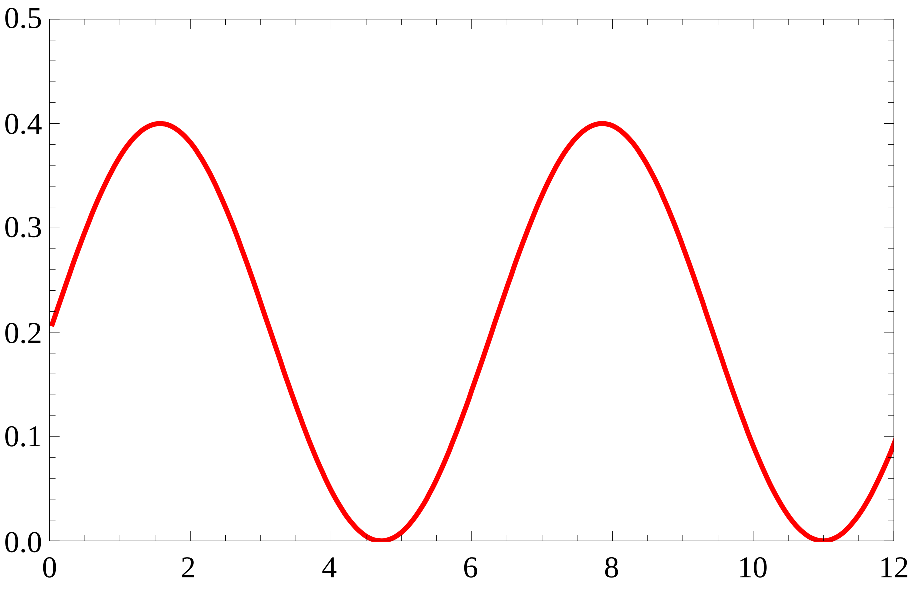
- Efficiency function for age a and service quota Δ:

$$J(a,0) = h(a), \quad J(a,\infty) = H(a)$$

$$J(a,\Delta) \triangleq \frac{P\{S - a \leq \Delta \mid S > a\}}{E[\min\{S-a,\Delta\} \mid S > a]} = \frac{\int_a^{a+\Delta} f(y)dy}{\int_a^{a+\Delta} \overline{F}(y)dy}$$

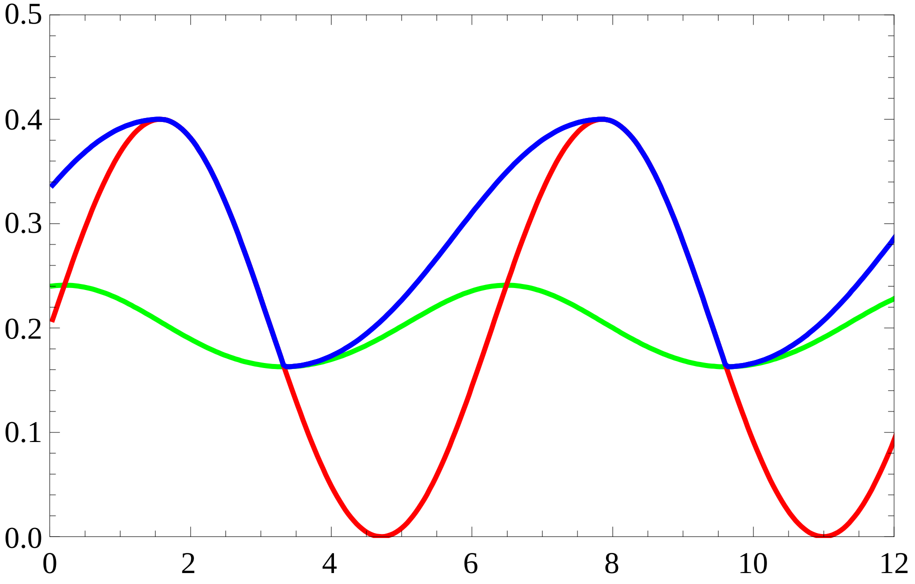
Example 6
Oscillating hazard rate

$$h(x) = \frac{1 + \sin x}{5}$$



Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)



Example 6
Oscillating hazard rate

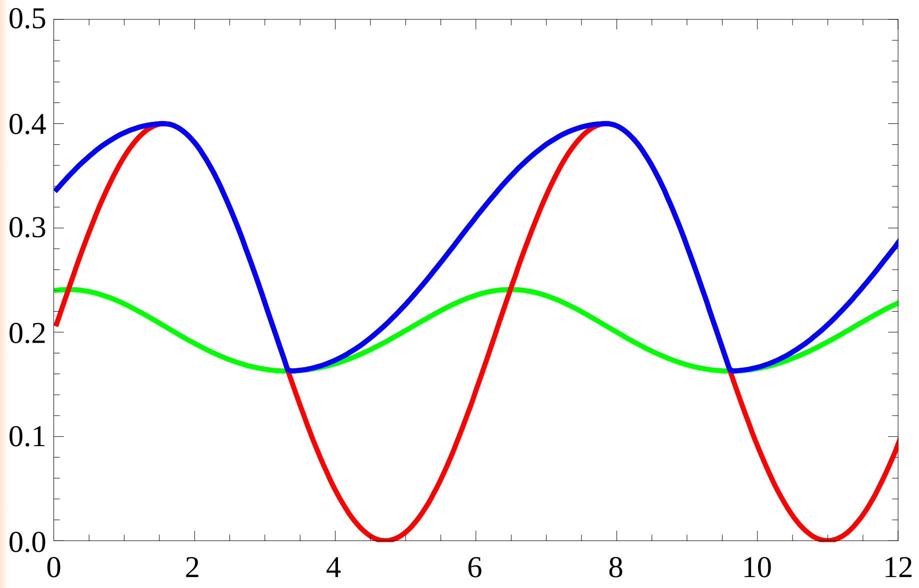
Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

NOTE!

h(x) continuous

$\Rightarrow$

G(x) continuous



Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

NOTE!

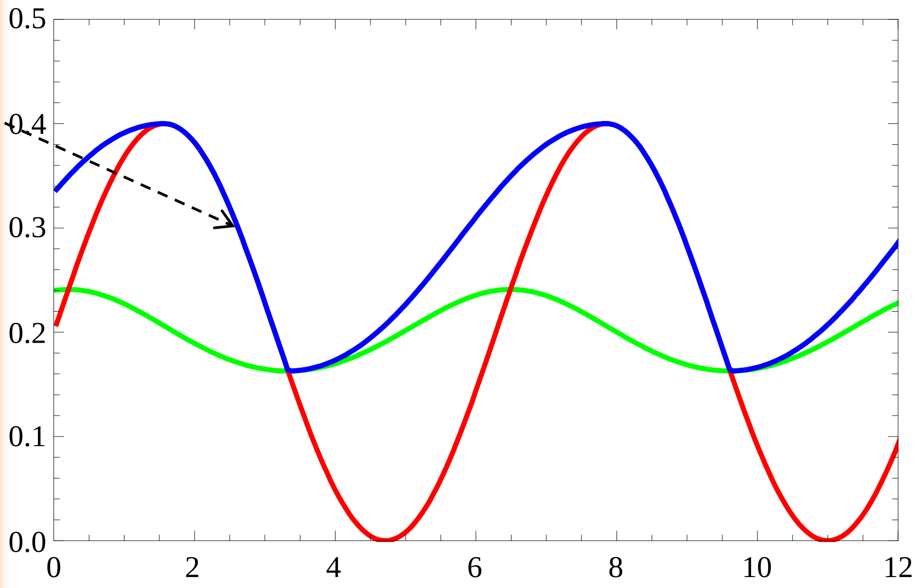
G(x) decreasing

⇒

h(x) decreasing

and

G(x) = h(x)



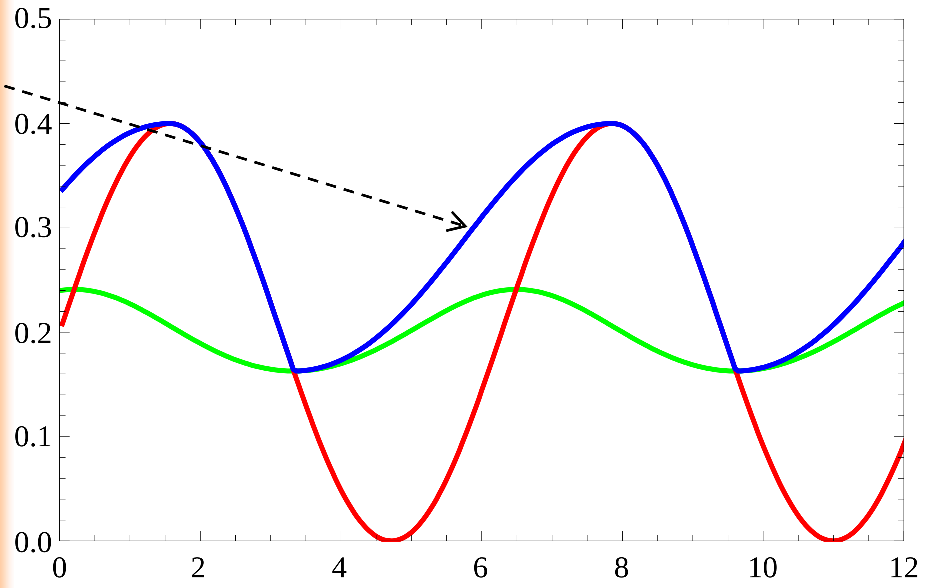
Aalto University
School of Science
and Technology

Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

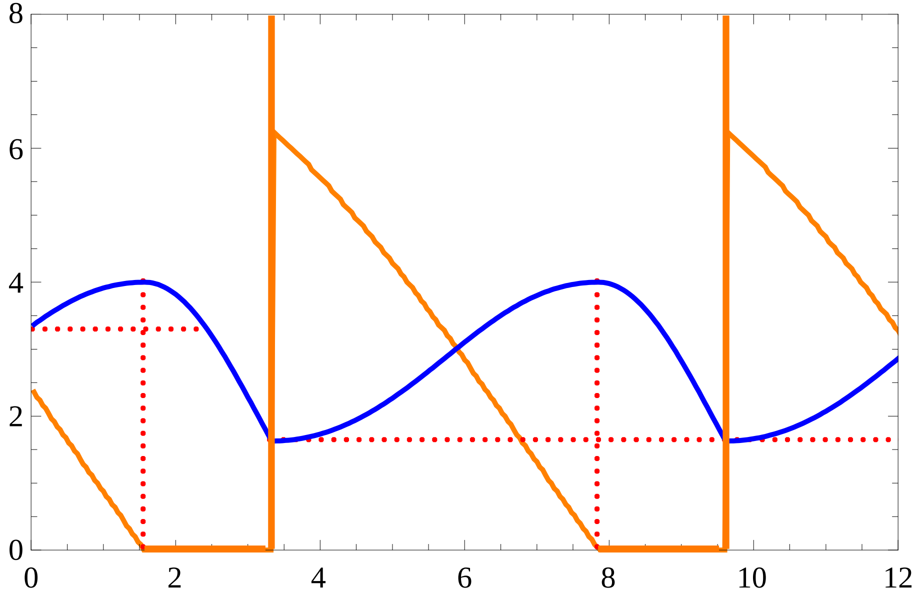
NOTE!

h(x) increasing
or
H(x) increasing
$\Rightarrow$
G(x) increasing



Example 6
Oscillating hazard rate

Gittins index G(x) (rescaled)
optimal service quota $\Delta^*(x)$



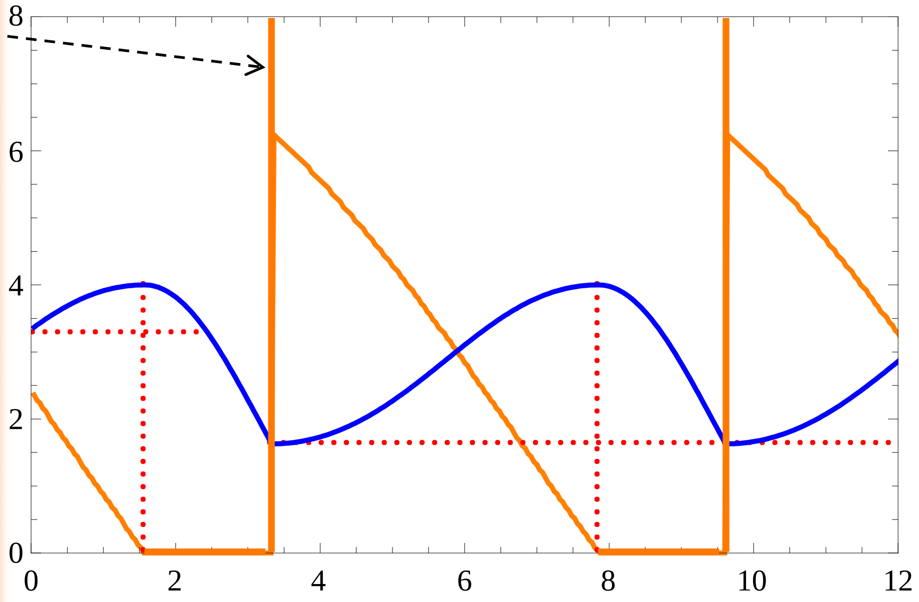
Example 6
Oscillating hazard rate

Gittins index G(x) (rescaled)

optimal service quota Δ*(x)

NOTE!

Here Δ*(x) = ∞



Aalto University
School of Science
and Technology

Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Continuity result

- Property:

$$f(x) \text{ is continuous for all } x$$
$$\Leftrightarrow h(x) \text{ is continuous for all } x$$
$$\Leftrightarrow J(x,d) \text{ is continuous for all } x,d$$

- Proposition:

$$h(x) \text{ is continuous for all } x$$
$$\Rightarrow G(x) \text{ is continuous for all } x$$

Monotonicity result 1

- Proposition:

$$h(x) \text{ strictly decreasing for all } x \in (a,b)$$

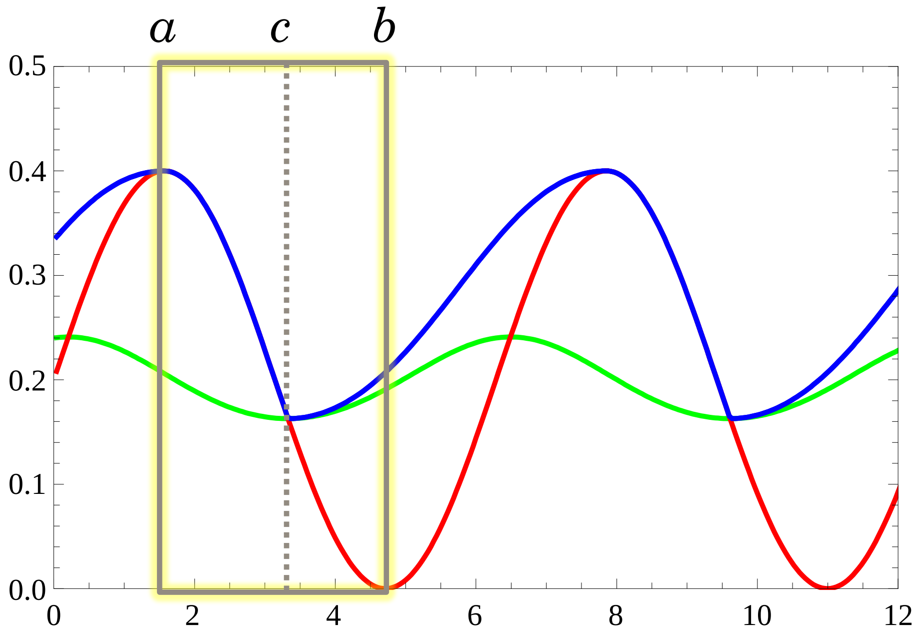
$$\Rightarrow$$

$$G(x) \text{ strictly decreasing for all } x \in (a,c),$$

$$G(x) \text{ increasing for all } x \in (c,b)$$

Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)



Monotonicity result 2

- Proposition:

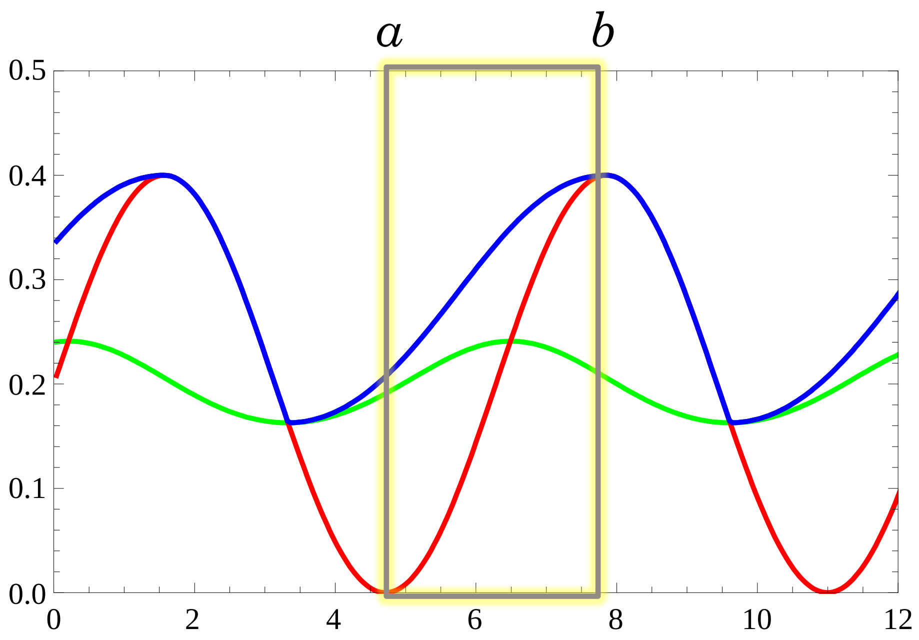
$$h(x) \text{ increasing for all } x \in (a,b)$$

$$\Rightarrow$$

$$G(x) \text{ increasing for all } x \in (a,b)$$

Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)



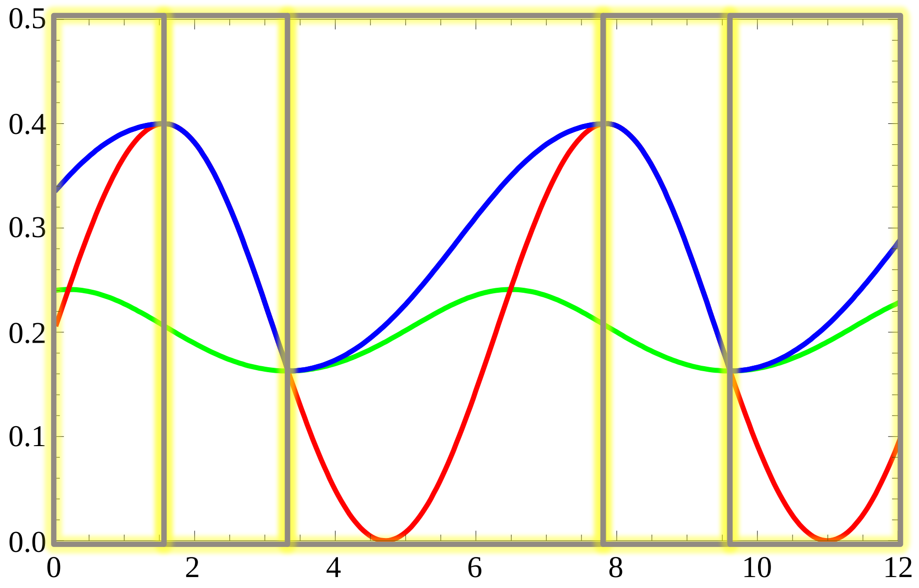
Continuity and monotonicity result

- Summary:

$$h(x) \text{ is continuous and piecewise monotonic for all } x$$
$$\Rightarrow G(x) \text{ is continuous and piecewise monotonic for all } x$$

Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

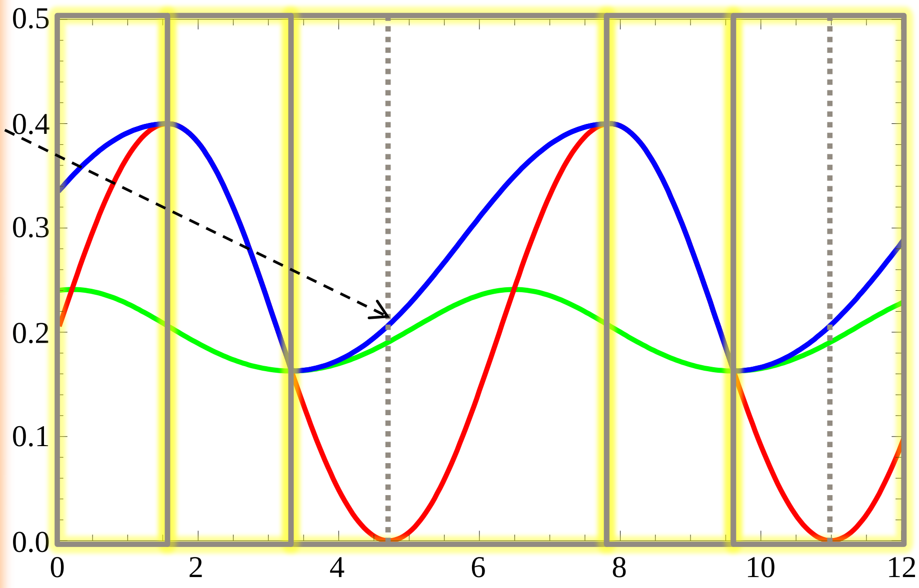


Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

NOTE!

Continuity needed here



Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Monotonicity in finite intervals 1

- Proposition:

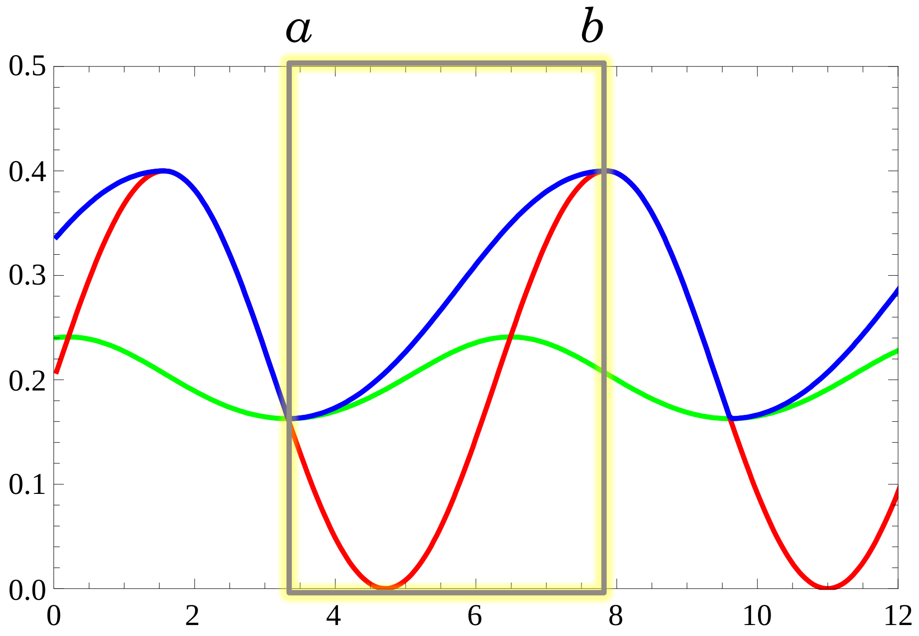
$$G(x) \text{ is strictly increasing for all } x \in (a,b)$$

$$\Leftrightarrow$$

$$G(x) > h(x) \text{ for all } x \in (a,b)$$

Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)



Monotonicity in finite intervals 2

- Proposition:

$$G(x) \text{ is increasing for all } x \in (a,b)$$

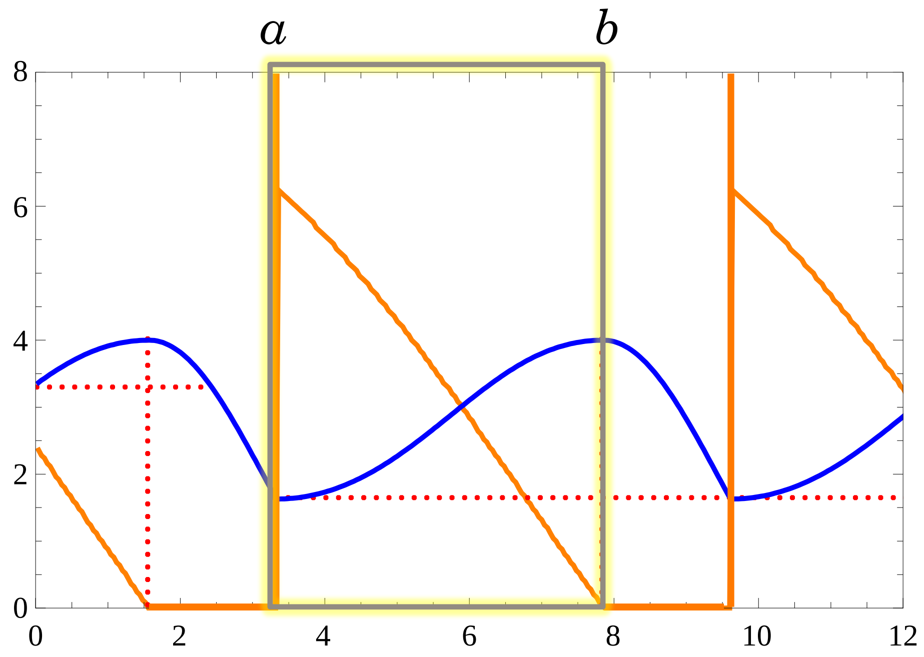
$$\Leftrightarrow$$

$$\Delta^*(x) > 0 \text{ for all } x \in (a,b)$$

Example 6
Oscillating hazard rate

Gittins index G(x) (rescaled)

optimal service quota $\Delta^*(x)$



Monotonicity in finite intervals 3

- Proposition:

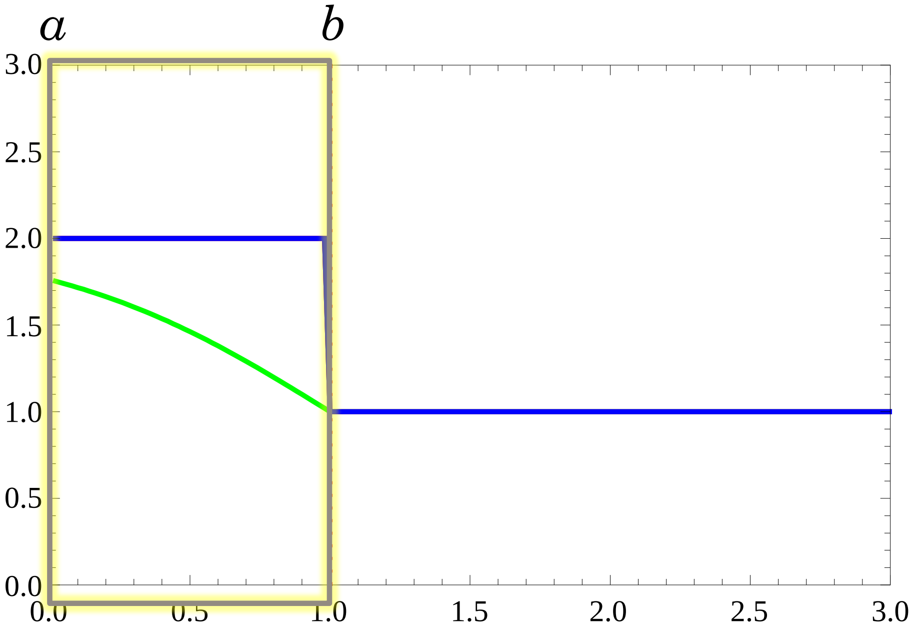
$$G(x) \text{ is constant for all } x \in (a,b)$$

$$\Leftrightarrow$$

$$G(x) = h(x) \;\text{ and }\; \Delta^{*}(x) > 0 \;\text{ for all } x \in (a,b)$$

Example 3
Decreasing hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)



Aalto University
School of Science
and Technology

Monotonicity in finite intervals 4

- Proposition:

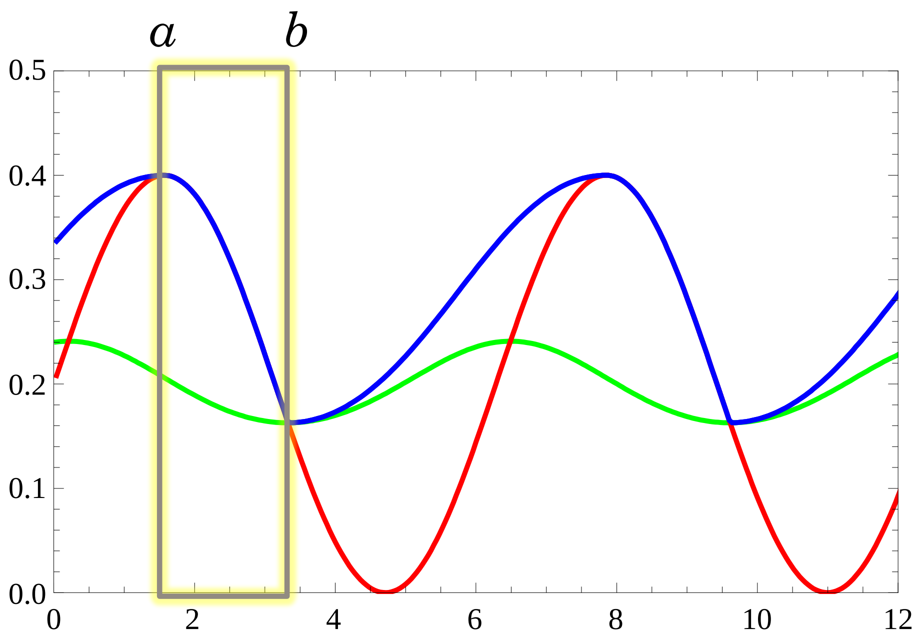
$$G(x) \text{ is decreasing for all } x \in (a,b)$$

$$\Leftrightarrow$$

$$G(x) = h(x) \text{ for all } x \in (a,b)$$

Example 6
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)



Monotonicity in finite intervals 5

- Proposition:

$$G(x) \text{ is strictly decreasing for all } x \in (a,b)$$

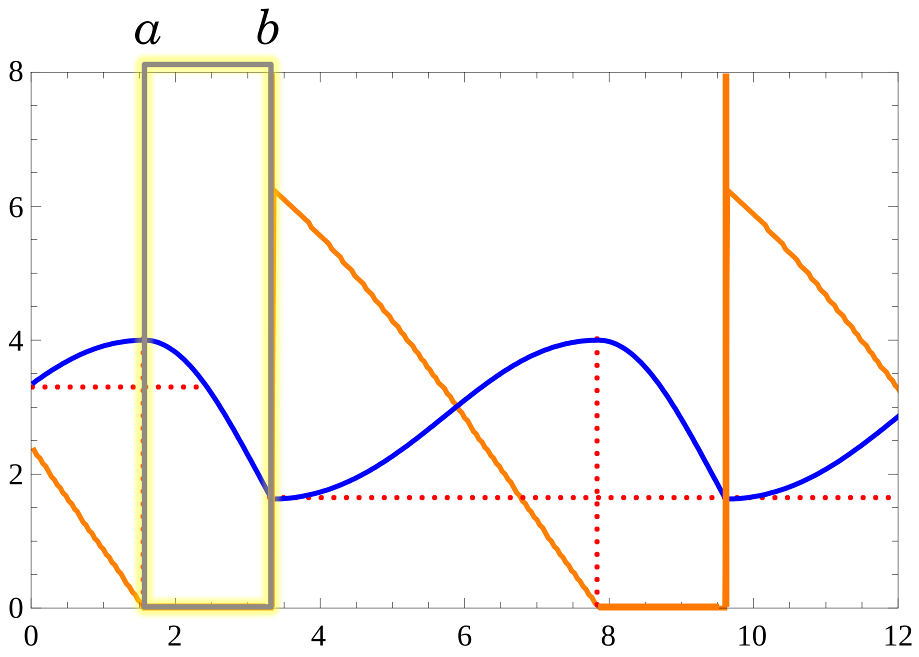
$$\Leftrightarrow$$

$$\Delta^{*}(x) = 0 \text{ for all } x \in (a,b)$$

Example 6
Oscillating hazard rate

Gittins index G(x) (rescaled)

optimal service quota $\Delta^*(x)$



Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Monotonicity in infinite intervals 1

- Proposition:

$$G(x) \geq G(a) \text{ for all } x \in (a, \infty)$$

$$\Leftrightarrow$$

$$H(x) \geq H(a) \text{ for all } x \in (a, \infty)$$

$$\Leftrightarrow$$

$$G(a) = H(a)$$

Example 6
Oscillating hazard rate

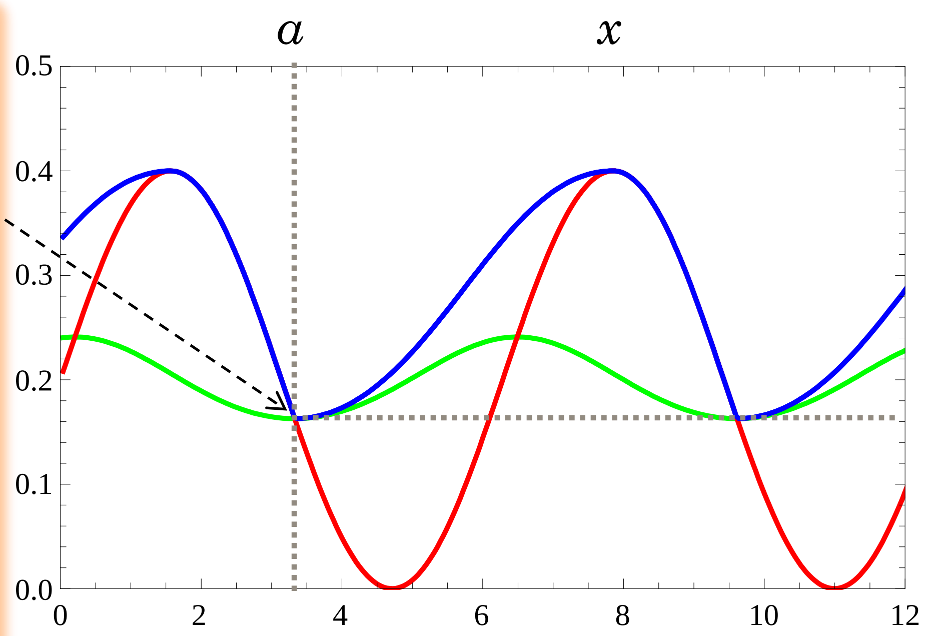
Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

NOTE!

G(a) = H(a)

G(x) ≥ G(a)

H(x) ≥ H(a)



Monotonicity in infinite intervals 2

- Proposition:

$$G(x)\text{ is increasing for all }x\in(a,\infty)$$

$$\Leftrightarrow$$

$$H(x)\text{ is increasing for all }x\in(a,\infty)$$

$$\Leftrightarrow$$

$$G(x)=H(x)\ \text{ for all }x\in(a,\infty)$$

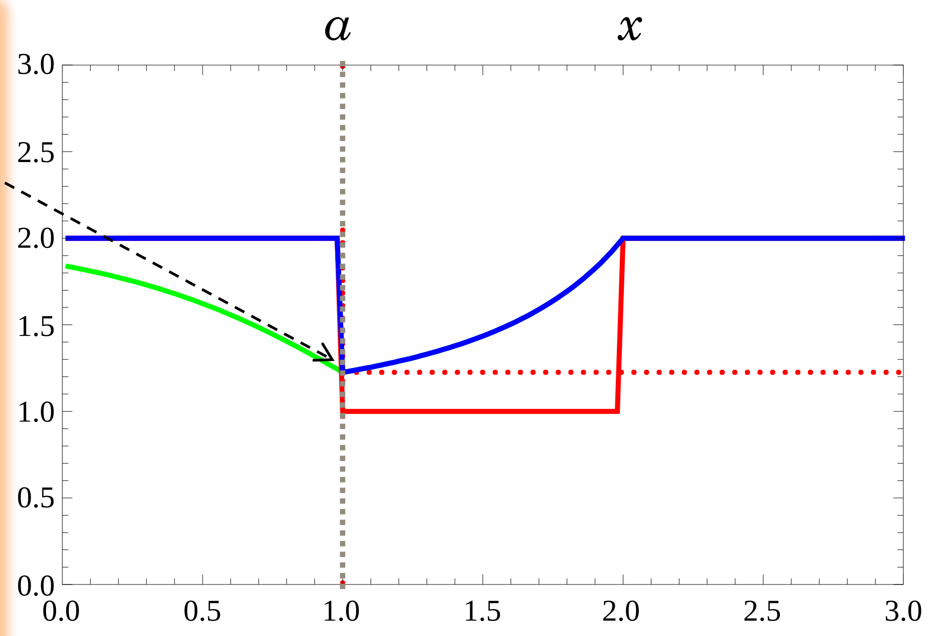
Example 5
Decreasing-increasing hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

NOTE!

G(x) = H(x)

for all x > a



Monotonicity in infinite intervals 3

- Proposition:

$$G(x) \text{ is decreasing for all } x \in (a,\infty)$$

$$\Leftrightarrow$$

$$h(x) \text{ is decreasing for all } x \in (a,\infty)$$

$$\Leftrightarrow$$

$$G(x) = h(x) \quad \text{for all } x \in (a,\infty)$$

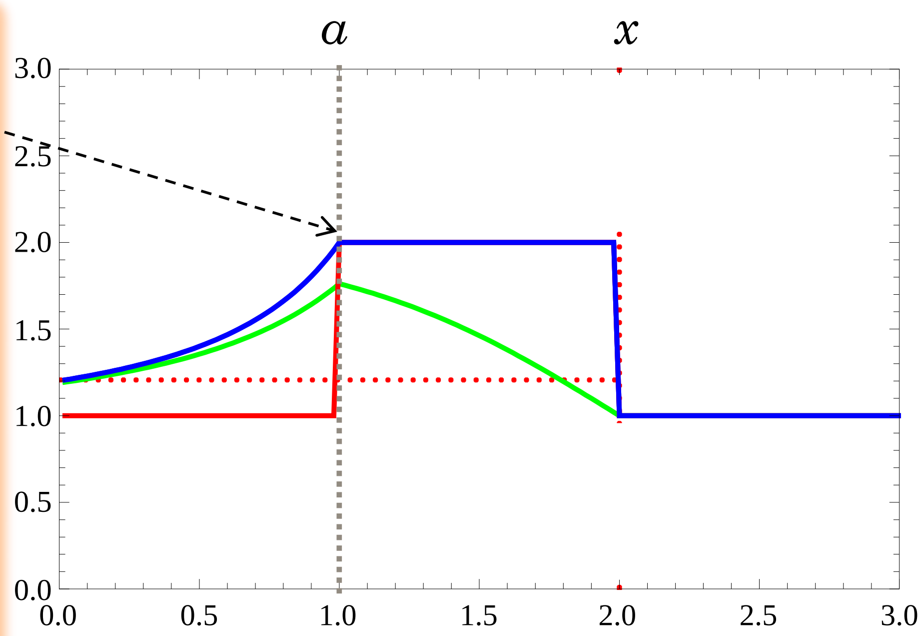
Example 4
Increasing-decreasing hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)

NOTE!

G(x) = h(x)

for all x > a



Monotonicity in infinite intervals 4

- Proposition:

$$G(x) \text{ is constant for all } x \in (a,\infty)$$

$$\Leftrightarrow$$

$$H(x) \text{ is constant for all } x \in (a,\infty)$$

$$\Leftrightarrow$$

$$h(x) \text{ is constant for all } x \in (a,\infty)$$

$$\Leftrightarrow$$

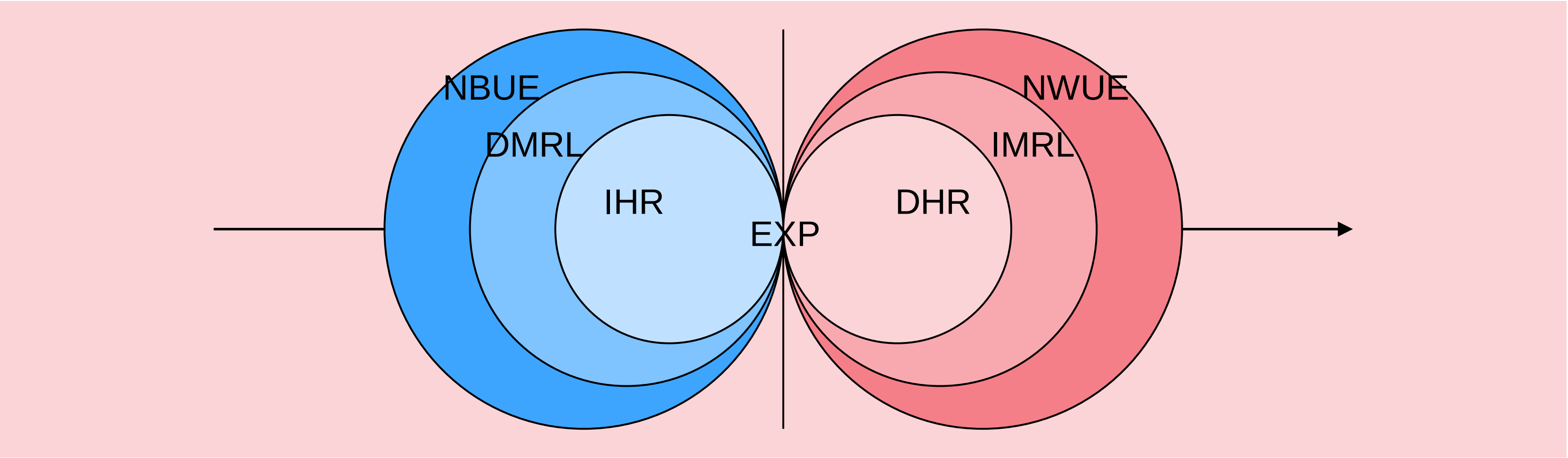
$$G(x) = H(x) = h(x) \ \text{ for all } x \in (a,\infty)$$

Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Service time distribution classes

- Service times are
  - IHR [DHR] if $h(x)$ is increasing [decreasing]
  - DMRL [IMRL] if $H(x)$ is increasing [decreasing]
  - NBUE [NWUE] if $H(0) \leq [\geq] H(x)$



NBUE service times

- Corollary:

$$G(x) \geq G(0) \text{ for all } x$$

$$\Leftrightarrow$$

$$\text{Service times are NBUE}$$

$$\Leftrightarrow$$

$$G(0) = H(0)$$

DMRL service times

- Corollary:

$$G(x) \text{ is increasing for all } x$$

$$\Leftrightarrow$$

$$\text{Service times are DMRL}$$

$$\Leftrightarrow$$

$$G(x) = H(x) \quad \text{for all } x$$

DHR service times

- Corollary:

$$G(x) \text{ is decreasing for all } x$$

$$\Leftrightarrow$$

$$\text{Service times are DHR}$$

$$\Leftrightarrow$$

$$G(x) = h(x) \quad \text{for all } x$$

EXP service times

- Corollary:

$$G(x) \text{ is constant for all } x$$

$$\Leftrightarrow$$

$$\text{Service times are EXP}$$

$$\Leftrightarrow$$

$$G(x) = H(x) = h(x) \text{ for all } x$$

Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Gittins index policy

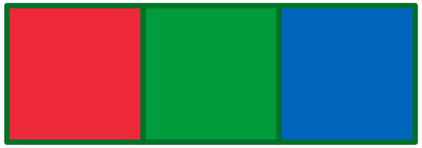
- Definition:
  - Gittins index policy (GI) gives service to the job i with the highest Gittins index $G(a_i)$.
- Known result:
  - Any GI is optimal for any service time distributions
- Observations:
  - Any MAS is a GI if and only if $G(a) \geq G(0)$ for all a.
  - LAS is a GI if and only if $G(a)$ is decreasing for all a.

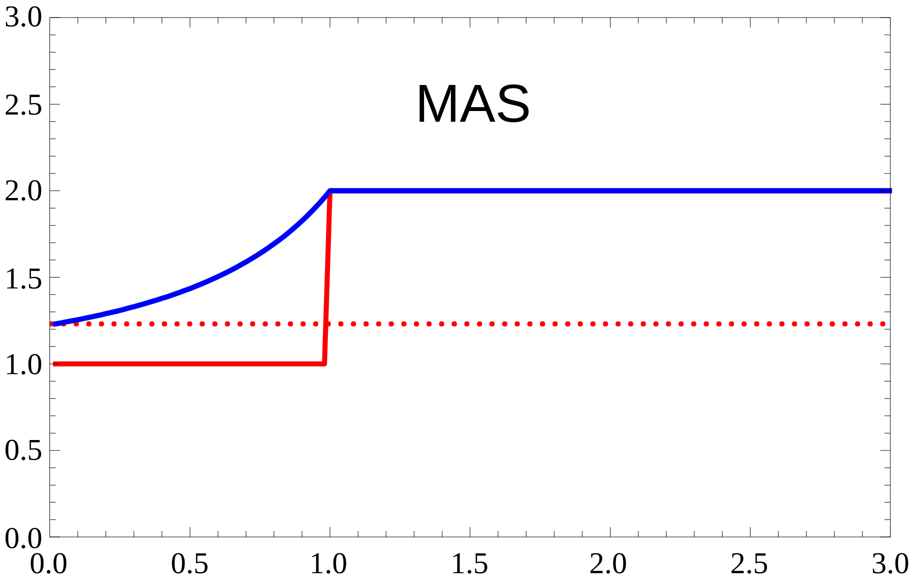
NBUE service times

- Corollary:

$$\text{Any MAS is optimal} \Leftrightarrow \text{Service times are NBUE}$$

Example 2
Increasing hazard rate



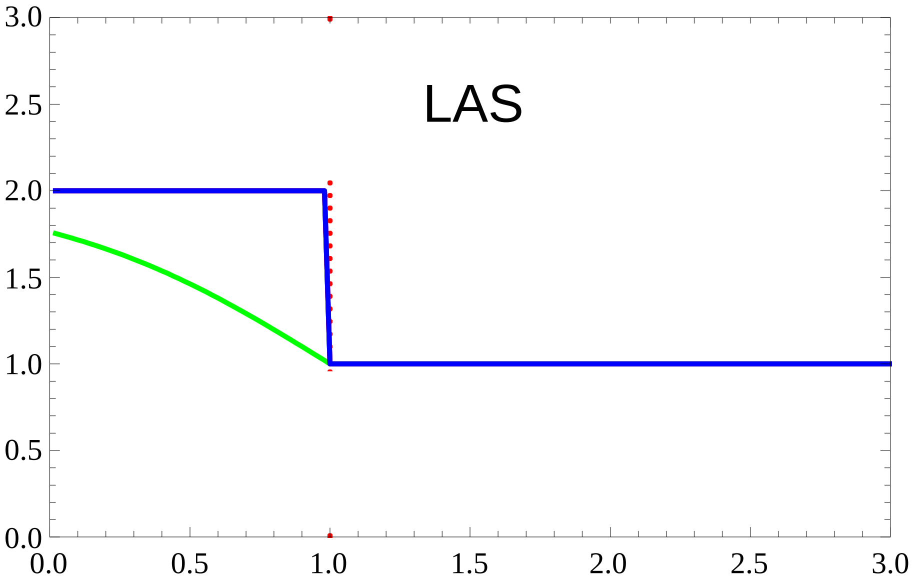


DHR service times

- Corollary:

$$\text{LAS is optimal} \iff \text{Service times are DHR}$$

Example 3
Decreasing hazard rate



NBUE+DHR service times

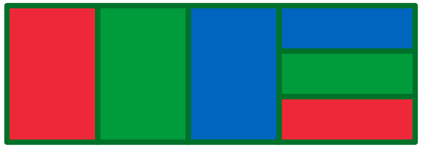
- Corollary:

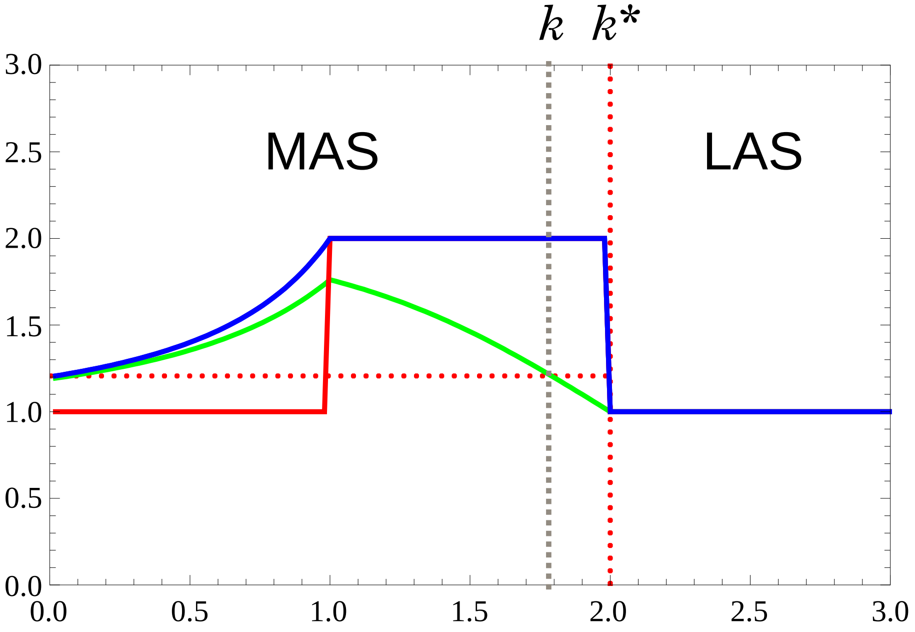
$$\text{Service times are NBUE} + \text{DHR}(k)$$

$$\Rightarrow$$

$$\text{MAS} + \text{LAS}(k^*) \text{ is optimal}$$

Example 4
Increasing-decreasing hazard rate



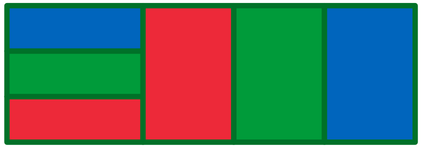


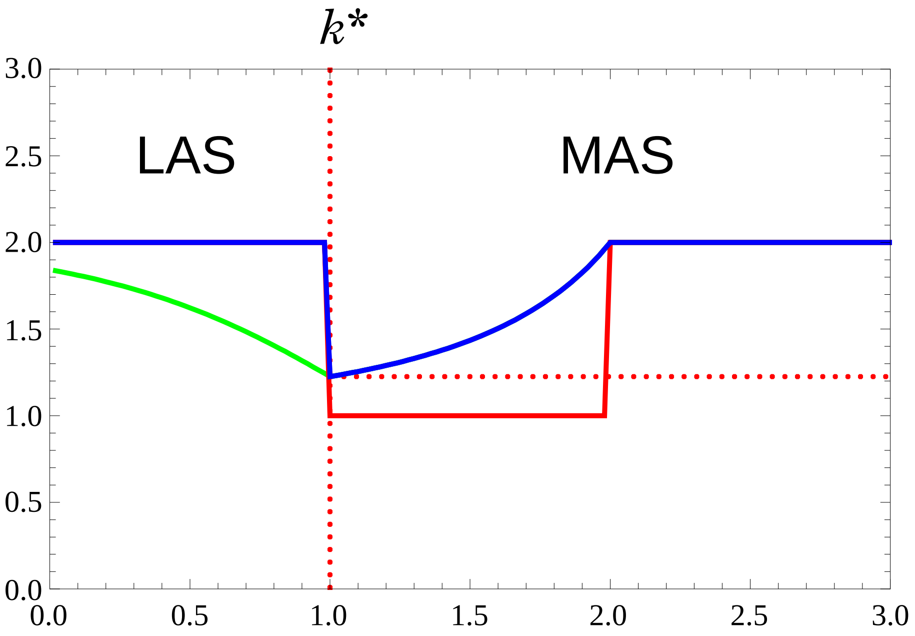
DHR+IHR service times

- Corollary:

$$\text{Service times are DHR} + \text{IHR}(k),$$
$$h(0) \geq H(\infty)$$
$$\Rightarrow$$
$$\text{LAS} + \text{MAS}(k^*) \text{ is optimal}$$

Example 5
Decreasing-increasing hazard rate





Transient system 1

- Assume h(x) is continuous and piecewise monotonic
- Corollary:

$$\text{Hazard rate } h(x) \text{ is first increasing}$$

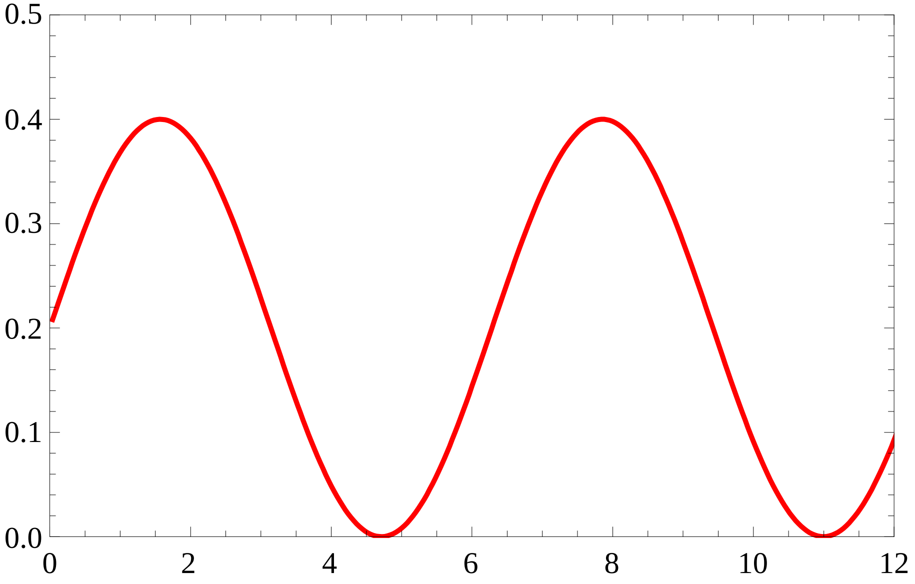
$$\Rightarrow$$

$$\text{MAS} + \text{LAS} + \text{MAS} + \ldots(k_1{*}, k_2{*}, \ldots) \text{ is optimal}$$

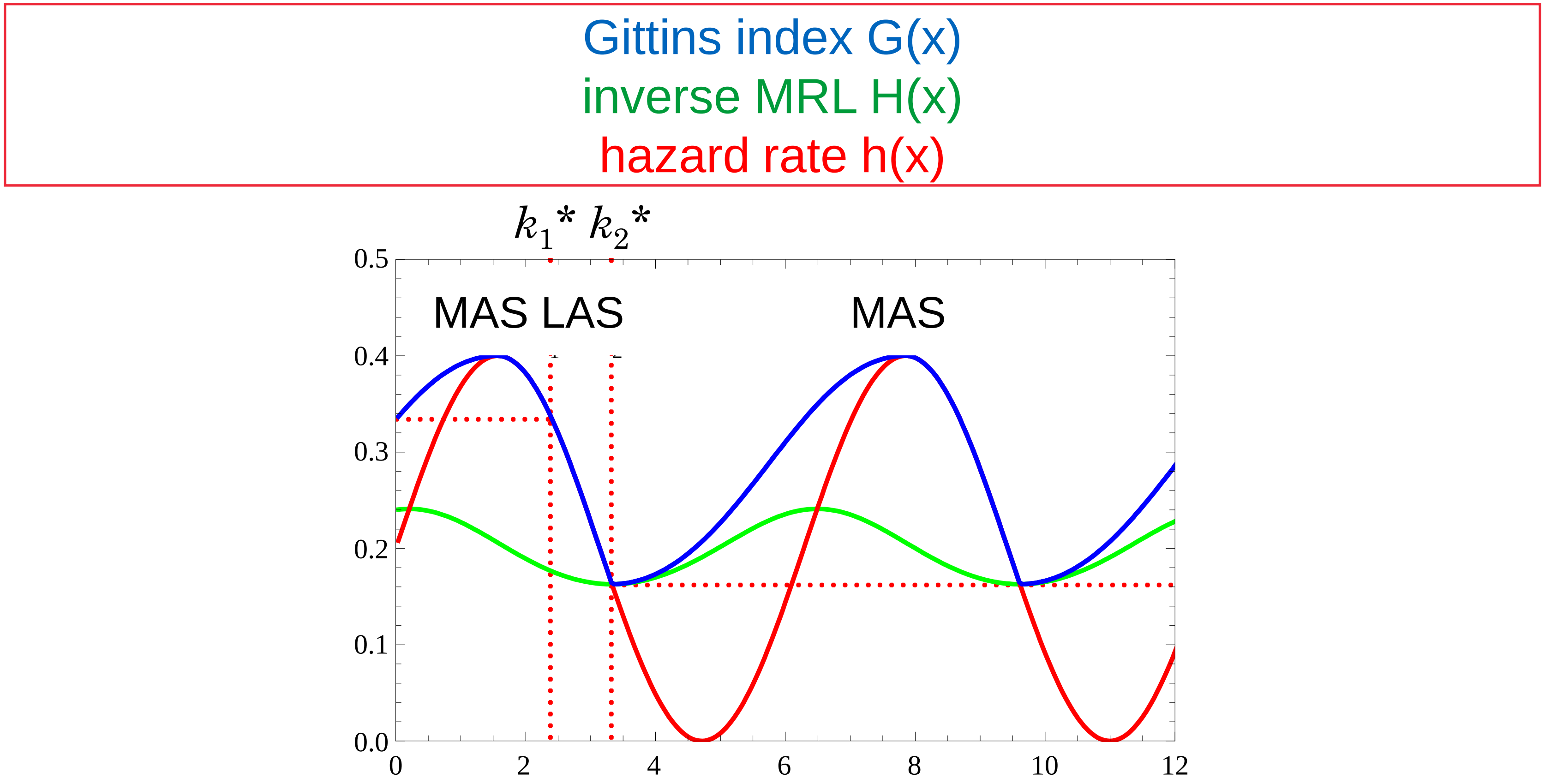
- MAS+LAS+MAS+… belongs to MLPS (Multi-Level Processor Sharing) policies, cf. Kleinrock (1976)

Example 6
Oscillating hazard rate

$$h(x) = \frac{1 + \sin x}{5}$$



Example 6
Oscillating hazard rate



Transient system 2

- Assume h(x) is continuous and piecewise monotonic
- Corollary:

$$\text{Hazard rate } h(x) \text{ is first decreasing}$$

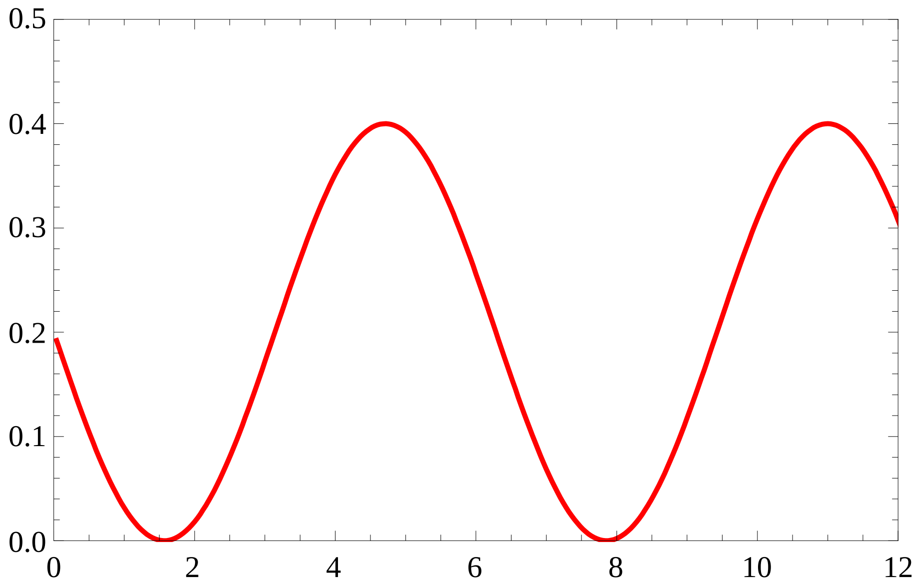
$$\Rightarrow$$

$$\text{LAS} + \text{MAS} + \text{LAS} + \ldots (k_1{*}, k_2{*}, \ldots) \text{ is optimal}$$

- LAS+MAS+LAS+… belongs to MLPS (Multi-Level Processor Sharing) policies, cf. Kleinrock (1976)

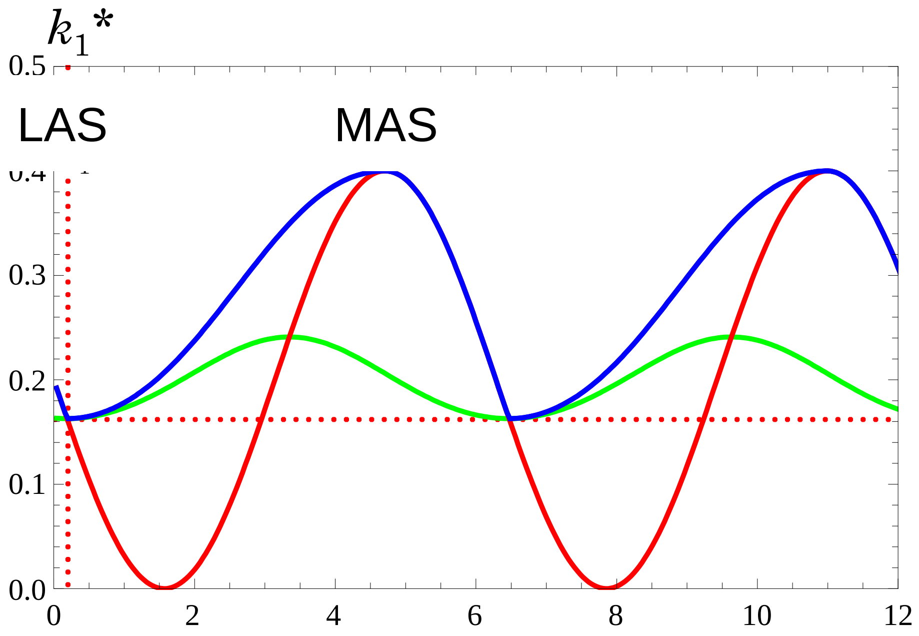
Example 7
Oscillating hazard rate

$$h(x) = \frac{1 - \sin x}{5}$$



Example 7
Oscillating hazard rate

Gittins index G(x)
inverse MRL H(x)
hazard rate h(x)



Outline

- Introduction
- Gittins index for single-server queues
- Continuity and monotonicity result
- Monotonicity in finite intervals
- Monotonicity in infinite intervals
- Service time distribution classes
- Gittins index policy
- Summary

Additional reading

- S. Aalto, U. Ayesta and R. Righter,
  On the Gittins index in the M/G/1 queue,
  Queueing Systems 63, 437-458, 2009

- S. Aalto, U. Ayesta and R. Righter,
  Properties of the Gittins index with application
  to optimal scheduling, submitted, 2010