Optimal scheduling discipline in a single-server queue with Pareto type service times

Samuli Aalto (Helsinki University of Technology)
Urtzi Ayesta (LAAS-CNRS)

Scheduling in an M/G/1 Queue

– Poisson arrivals with rate λ. Service requirements are i.i.d. with distribution $\underline{F}(x)=P[X \leq x]$. Complementary cumulative distribution denoted by $\overline{F}(x) = 1 - F(x)$

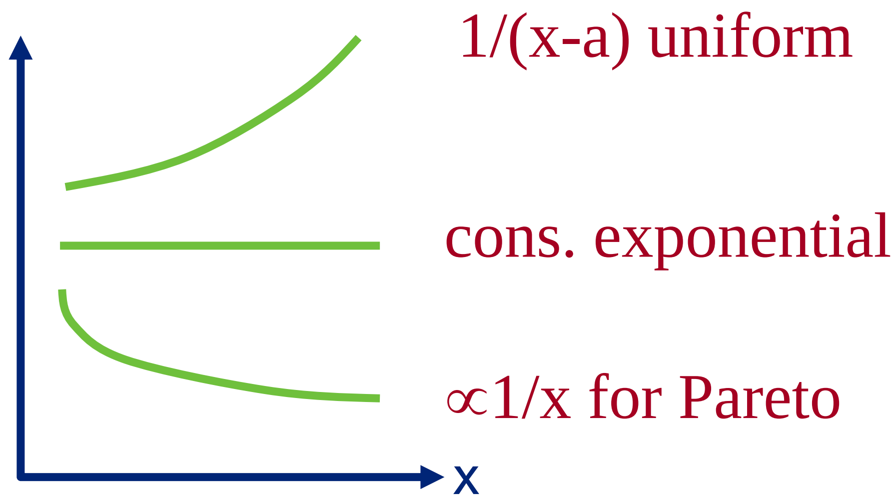
– Attained service is known (total service requirement unknown)

– Optimality criterion: Mean number of jobs in the system

Monotonous Hazard Rate

– Hazard rate of a distribution function: h(x)dx=P[x< X ≤ x+dx | X > x]

$$h(x)=\frac{f(x)}{1-F(x)}$$



– IHR: Non-preemptive discipline (FCFS etc.)

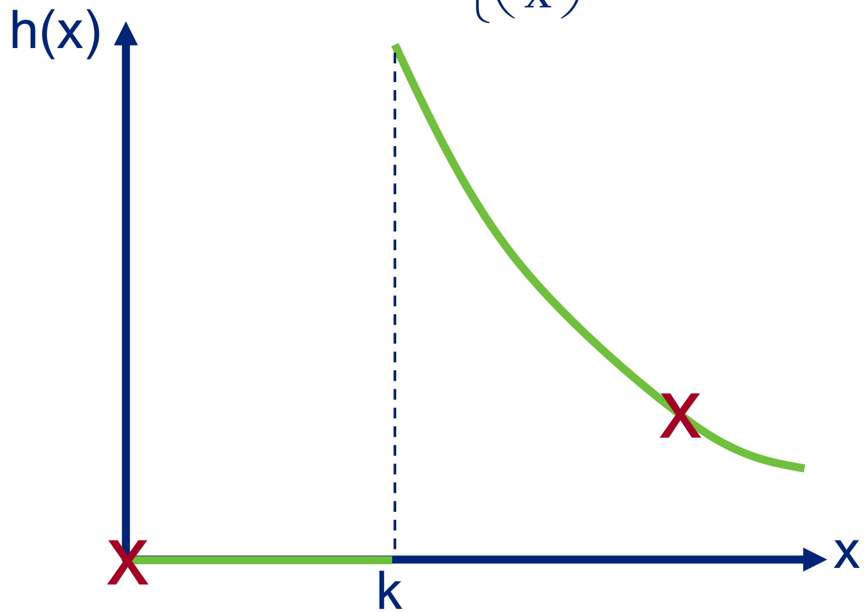
– Exponential: M/M/1 Mean number of jobs is policy independent

– DHR: Least Attained Service (LAS) is optimal. The job(s) who has attained the least amount of service is served.

Which scheduling when HR not monotonous?

- Assume that the distribution is defined in an interval [k,∞)? for example a Pareto-type distribution

$$\bar{F}(x) = \begin{cases} 1, & 0 < x \leq k \\ \left(\dfrac{k}{x}\right)^{\alpha}, & x > k \end{cases}$$



- If the support is bounded, that is, if F(x)=1 for all p≤x ?

Optimal discipline for general service requirements

– Gittins' index policy.
  - To each job present in the system, assign an index equal to

$$G(a) = \sup_{\Delta \geq 0} J(a, \Delta) \quad \text{where} \quad J(a, \Delta) = \frac{\int_0^{\Delta} f(a+\Delta)}{\int_0^{\Delta} \overline{F}(a+\Delta)}$$

  - Pick the job with highest index value, and assign him a service quota $\Delta^*(a)$

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

  - Another job will start being served when:
    - The previously selected job receives $\Delta^*(a)$ units of service
    - The previously selected job departs from the system
    - A new job arrives to the queue

Gittins optimal policy

- Introduced by Sevcik [1974] for static scheduling. Optimality in Stochastic setting by Gittins [1989].

- Theorem [Gittins]. The index policy minimizes the mean number of jobs in the system among all non-anticipating scheduling policies

Properties of Gittins

- **Theorem:** For all $a \le x \le a+\Delta^*(a)$,
  - $G(x) \ge G(a)$
  - $x+\Delta^*(x) \le a+\Delta^*(a)$

Sketch of the proof: Take $a=0$ and let $\Delta^*(0)=\text{argmax}_\Delta \{J(0, \Delta)\}$. For all $0 \le x \le \Delta^*(0)$, there exists a function $p(x) \le 1$ such that

$$J(0,\Delta^*(0))=p(x)\, J(0, x) + (1-p(x))\, J(x,\Delta^*(0)-x).$$

But $J(0,\Delta^*(0)) \ge J(0,x)$, thus $J(x,\Delta^*(0)-x) \ge J(0,\Delta^*(0))$.

Now it follows that

$$G(x) \ge J(x,\Delta^*(0)-x) \ge J(0,\Delta^*(0)) = G(0).$$

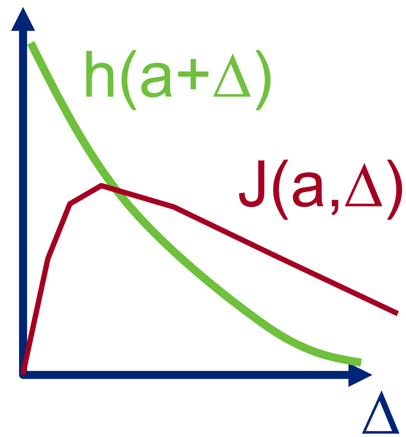
Gittins index policy

- Theorem: The scheduling discipline that at any time assigns an infinitesimal quota to the job with highest G(x) is equivalent (sample-pathwise) to the Gittins policy

- For non-anticipative disciplines, the hazard rate suffices to characterize the optimal scheduling discipline.

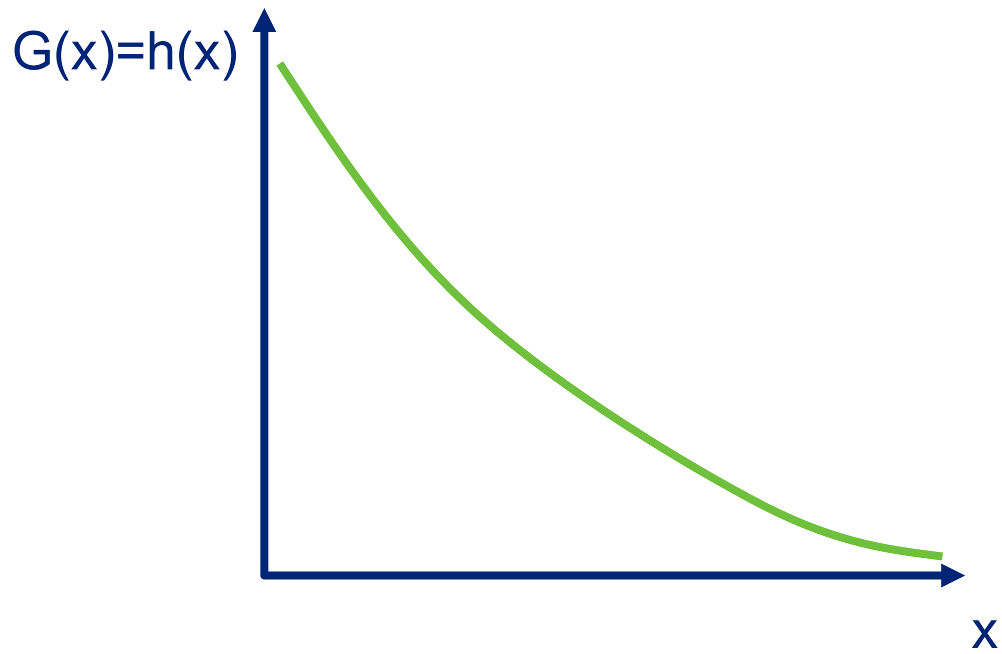
- Theorem: For any attained service $a \geq 0$,

$$G(a) = h\left(a + \Delta^*(a)\right)$$



Sketch of the proof: $\frac{\partial}{\partial \Delta} J(a, \Delta) = 0 \Rightarrow J\left(a, \Delta^*(a)\right) = h\left(a + \Delta^*(a)\right)$

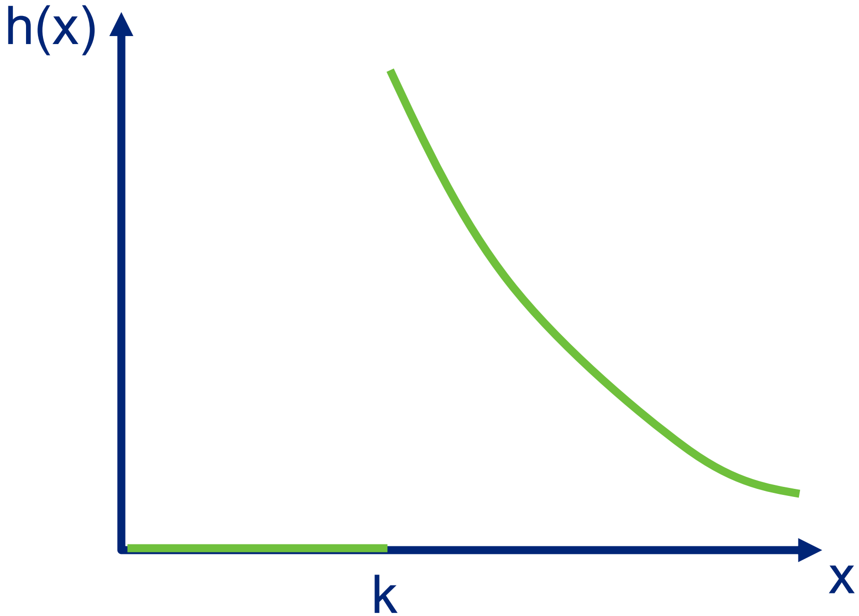
- Theorem: If the distribution is of type DHR, Least-Attained-Service minimizes the mean number of jobs in the system
- Sketch of the proof:
  - For any fixed a, J(a,Δ) is decresing with respect to Δ.
  - Then for all a, G(a)=J(a,0)=h(a), and note that h(a) is decreasing



- Similar result for IHR

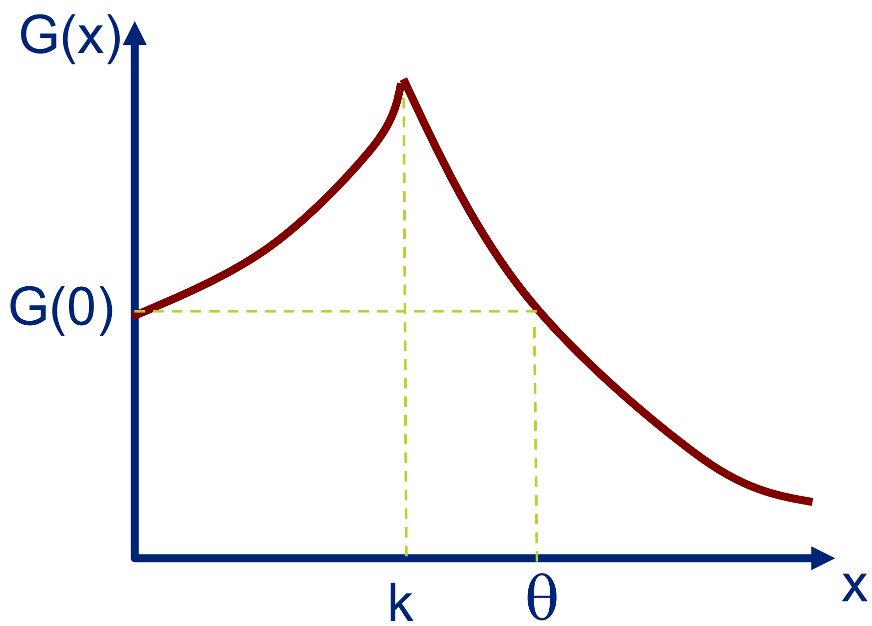
CDHR(k) or Pareto-type distributions

– 3 assumptions:
  - A1: $h(x)$ is constant for all $x < k$,
  - A2: $h(x)$ is decreasing for all $x \geq k$.
  - A3: $h(0) < h(k)$.



– Proposition. Assume that the service time distribution belongs to the class CDHR(k).

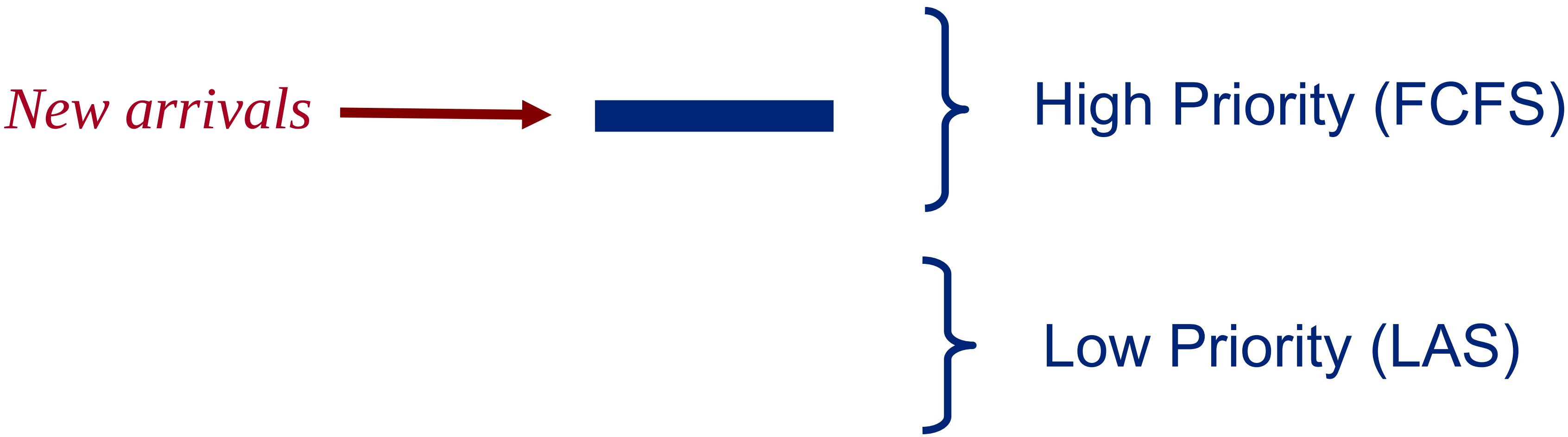
- (i) If assumption A3 is not satisfied, then $G(x)$ is decreasing for all x.
- (ii) If assumption A3 is satisfied, then,
  - $G(x) \geq G(0)$ for all $x < \theta$ and $\theta > k$,
  - $G(\theta) \leq G(0)$, and
  - $G(x)$ is decreasing for all $x \geq \theta$.



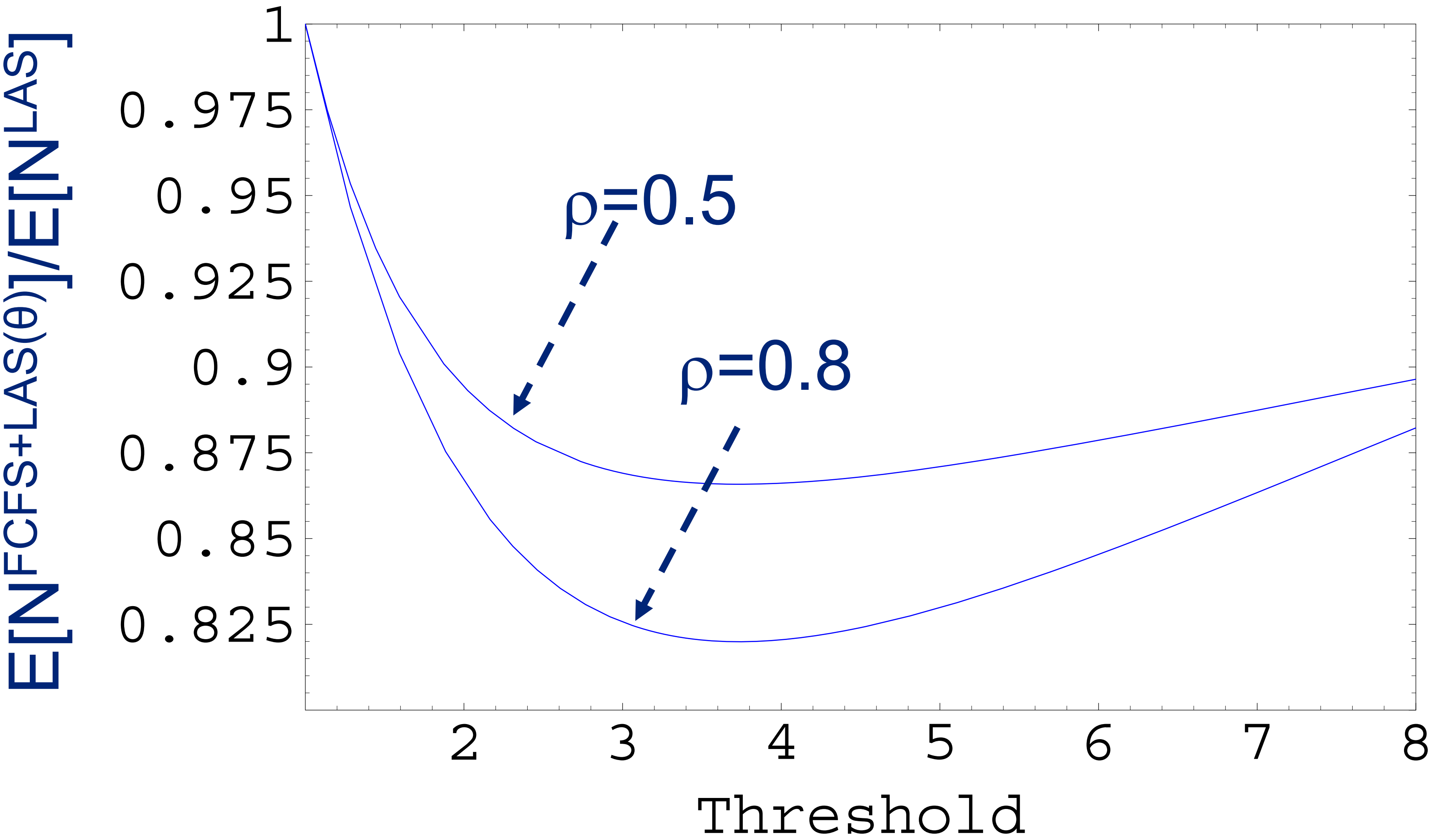
- Theorem: Assume that the service time distribution belongs to CDHR(k).
  - (i) If assumption A3 is not satisfied, then LAS is optimal.
  - (ii) If assumption A3 is satisfied, then there is $\theta > k$ such that FCFS+ LAS(θ) is optimal. The precise value of θ depends only on the parameters of the service time distribution.

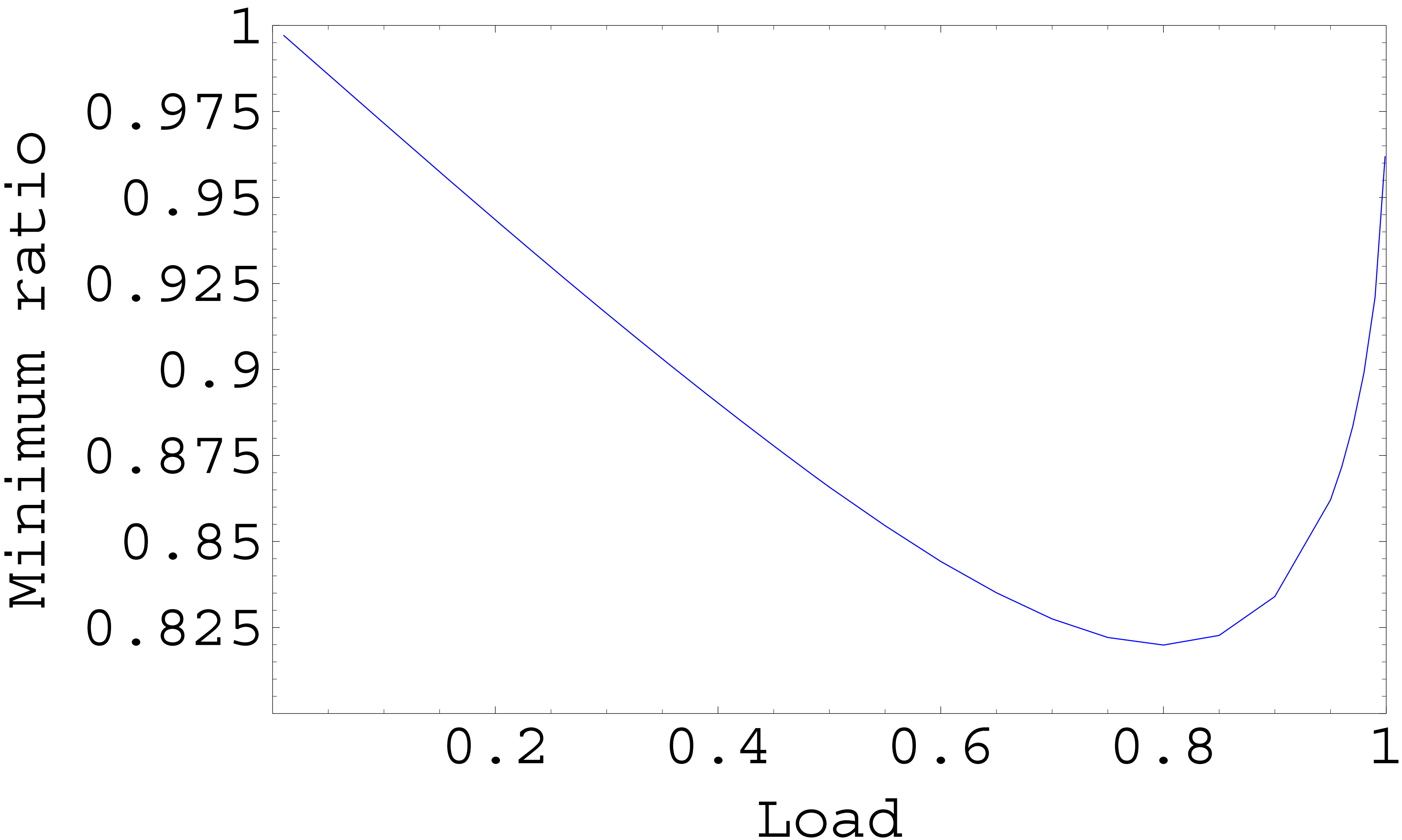
- FCFS+LAS(θ)
  - Classify jobs into two classes depending on the amount of attained service
    - High Priority: Jobs that have obtained less service than θ
    - Low Priority: Jobs that have obtained more service than θ

  - High Priority jobs served according to FCFS and Low Priority with LAS



Numerical example: Pareto distribution with k=1 and α =2

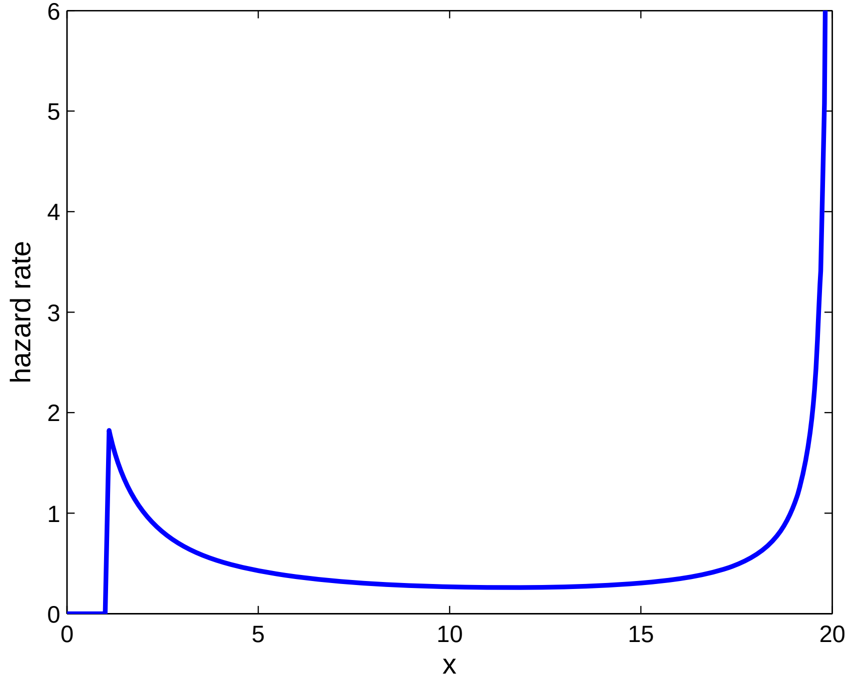




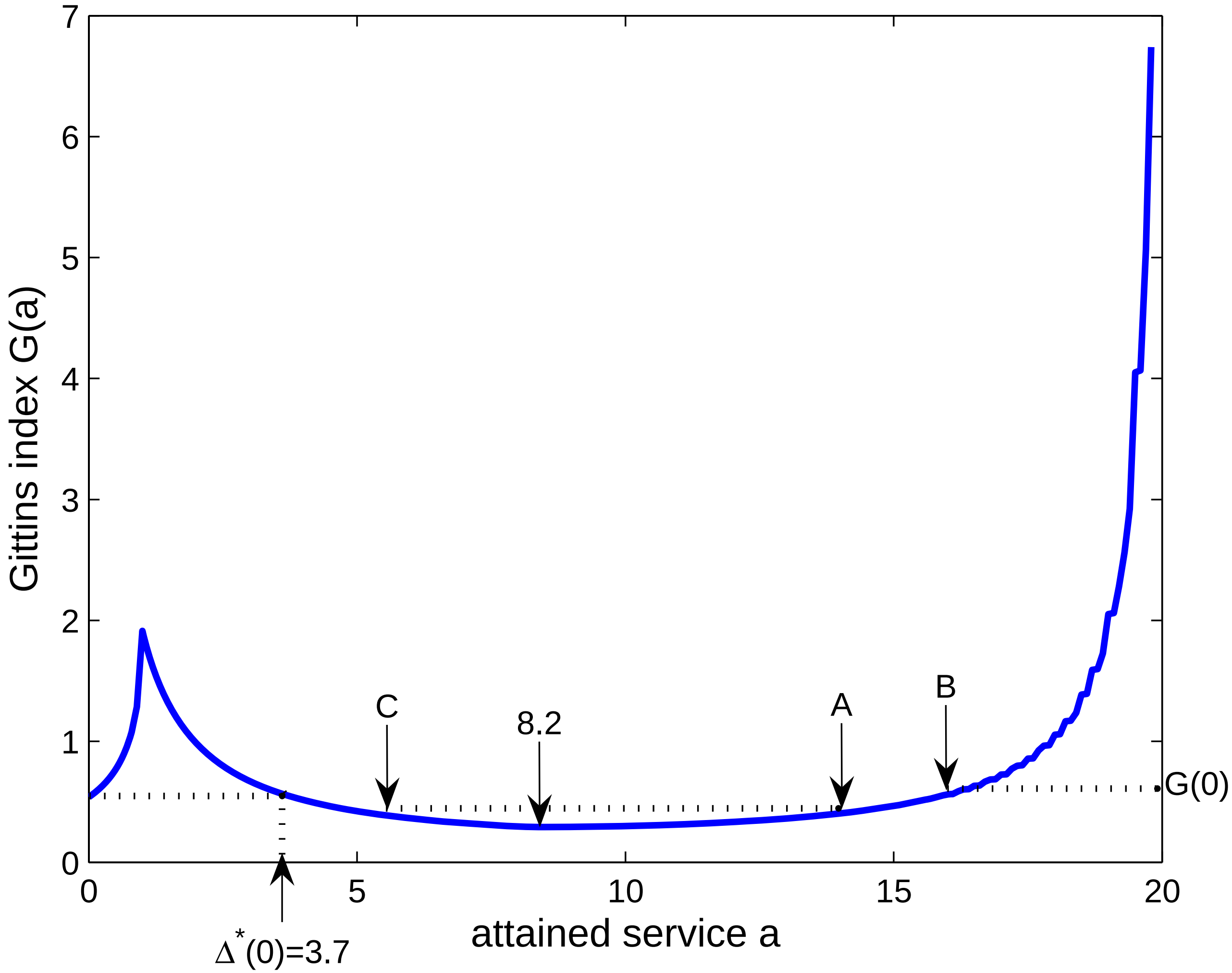
Minimum ratio
1
0.975
0.95
0.925
0.9
0.875
0.85
0.825
0.2
0.4
0.6
0.8
1
Load

Impact of an upper bound bounded distribution: Bounded Pareto

$$\overline{F}(x) = \begin{cases} 1, & 0 \le x < k, \\ 1 - \dfrac{1-(k/x)^\alpha}{1-(k/p)^\alpha}, & k \le x < p, \\ 0, & x \ge p. \end{cases} \qquad h(x) = \begin{cases} 0, & 0 \le x < k, \\ \dfrac{\alpha}{x\left(1-(x/p)^\alpha\right)}, & k \le x < p. \end{cases}$$



Gittins index for Bounded Pareto



Conclusion and future research

- In the set of non-anticipative disciplines, the hazard rate characterizes completely the optimal policy.
- Application of index policy for scheduling in multi-server systems?
  - How to cope with non work conserving property of networks?
- And with time-varying server capacity like in wireless systems?
- Scheduling in a G/G/1 queue. LAS and FCFS (with DHR and IHR respectively). What if hazard-rate is not monotone?
- Calculate performance metrics for a given function G(a).