Optimal Control in Call Centers

Benjamin Legros

Collaboration with Interact-iv.com

Ger Koole and Oualid Jouini

Outline

1. Markov Decision process Approach, a tool to find and prove (eventually) optimal policies in queueing models

    a. The approach

    b. The value iteration technique

    c. Limitations of the method, open questions, improvement.

2. Adaptive Routing

3. Routing decisions based on the waiting time

The rejection problem

Consider a single server queue with infinite buffer, an homogeneous Poisson arrival process with rate λ and an exponential server working with rate μ (M/M/1 queue).

A customer can be rejected at any time with cost *r* and a cost *c* is counted per customer waiting in the queue per time unit.

The objective is to minimize the long run expected cost of the system.

State definition: Number of customers in the system: *x*

Problem: The problem is a **continuous time** MDP, the uniformisation technique from Puterman (1994) allows to change this problem into a discrete one.

The rejection problem

We thus assume $\lambda + \mu = 1$ and define the value function:

Transitions:

$V_{n+1}(x) = c \times (x-1)^{+} + \lambda\, U_n(x+1) + \mu\, U_n(x-1)$,
for x>0,
$V_{n+1}(0) = \lambda\, U_n(1) + \mu\, U_n(0)$

Actions:

$$U_n(x) = \min(V_n(x), V_{n+1}(x-1) + r)$$

For each *n*, there is a minimizing action: **keep** or **reject**

The function from $\{0,1,\ldots,\mathrm{x},\ldots\} \rightarrow \{\text{keep}, \text{reject}\}$ is a **policy.**

The rejection problem

If it is optimal to reject in x, then it is optimal to reject in x+1 (threshold structure).

A necessary condition is

$$V(x+1) - V(x) - r \geq 0$$

induces

$$V(x+2) - V(x+1) - r \geq 0$$

So, if

V_n is convex then the condition is proven.

Generalization

From Puterman (1994) the property holds as n tends to infinity (convergence of V_{n+1}-V_n)

Usually, other monotonicity properties have the value function have to be proven (first order monotonicity).

In two dimensions other second order monotonicity results are often necessary conditions (generated by minimizing or maximizing actions):

$$V(x+1,y+1)+V(x,y) \geq V(x+1,y)+V(x,y+1)$$

and/or

$$V(x+2,y)+V(x,y+1) \geq V(x+1,y+1)+V(x+1,y)$$

Generalization/problems

1. Unbounded rates (abandonment for instance): truncation is a solution but the convexity is broken at the truncated state.
2. Multiserver case: to prove convexity the minimizing action forces to prove supermodularity, to prove supermodularity the minimizing action forces to prove second order supermodularity, to prove the second order supermodularity the service term imposes to prove concavity…

Generalization/problems

<u>3. Non exponential distribution:</u> the Coxian distribution can be used but the number of states increases.

<u>4. Non convex/concave performance measures</u>: an example is the percentile of the waiting time (80/20 rule).

<u>5. Non traditional state definition</u>: in call centers routing decisions are often taken based on the experienced waiting time

Unsolved problem

The slow server

Consider a model with a slow and a fast server. The objective is to minimize the time spent in the system.

The optimal policy is of threshold type. The slow server is used only when the queue size exceeds a threshold.

A long history….

Introduction of the problem

Krishnamoorthi, B. (1963). On poisson queue with two heterogeneous servers. *Operations research*, 11(3):321–330.

The proofs

Larsen, R. and Agrawala, A. (1983). Control of a heterogeneous two-server exponential queueing system. *Software Engineering, IEEE Transactions on*, (4):522–526.

Lin, W. and Kumar, P. (1984). Optimal control of a queueing system with two heterogeneous servers. *IEEE Transactions on Automatic Control*, 29(8):696–703.

Koole, G. (1995). A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems and Control Letters*, 26:301–303.

A long history….

With more than two servers….

Rykov, V. (2001). Monotone control of queueing systems with heterogeneous servers. *Queueing Systems*, 37:391–403.

Luh, H. P. and Viniotis, I. (2002). Threshold control policies for heterogeneous server systems. *Mathematical Methods of Operations Research*, 55(1):121–142.

But…

de Véricourt, F. and Zhou, Y. (2006). On the incomplete results for the heterogeneous server problem. *Queueing Systems*, 52:189–191.

So….

Mehrotra, V., Ross, K., Ryder, G., and Zhou, Y. (2012). Routing to manage resolution and waiting time in call centers with heterogeneous servers. *Manufacturing & service operations management*, 14(1):66–81.

Ozkan, E. and Kharoufeh, J. P. (2014). Optimal control of a two-server queueing system with failures. *Probability in the Engineering and Informational Sciences*, 28:489–527.

Rykov, V. V. and Efrosinin, D. V. (2009). On the slow server problem. *Automation and Remote Control*, 70(12):2013–2023.

Zhan, D. and Ward, A. (2014). Threshold routing to trade off waiting and call resolution in call centers. *Manufacturing & Service Operations Management*, 16(2):220–237.

Efrosinin, D. (2013). Queueing model of a hybrid channel with faster link subject to partial and complete failures. *Annals of Operations Research*, 202:75–102.

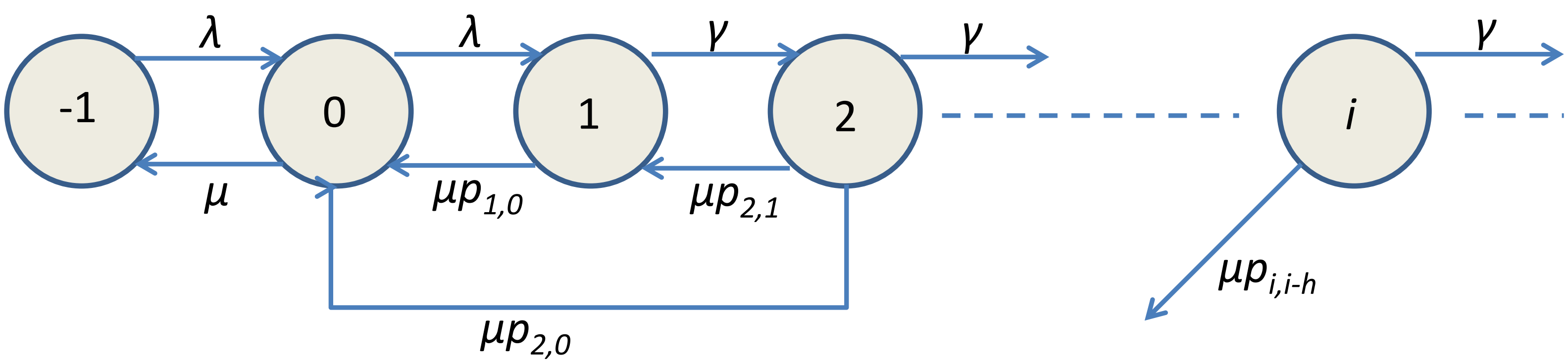
Cabral, F. (2005). The slow server problem for uninformed customers. *Queueing systems*, 50(4):353–370.

Armony, M. (2005). Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51(3-4):287–329.

Armony, M. and Ward, A. (2010). Fair dynamic routing in large-scale heterogeneous-server systems. *Operations Research*, 58(3):624–637.

2. Routing Decision on the Waiting Time

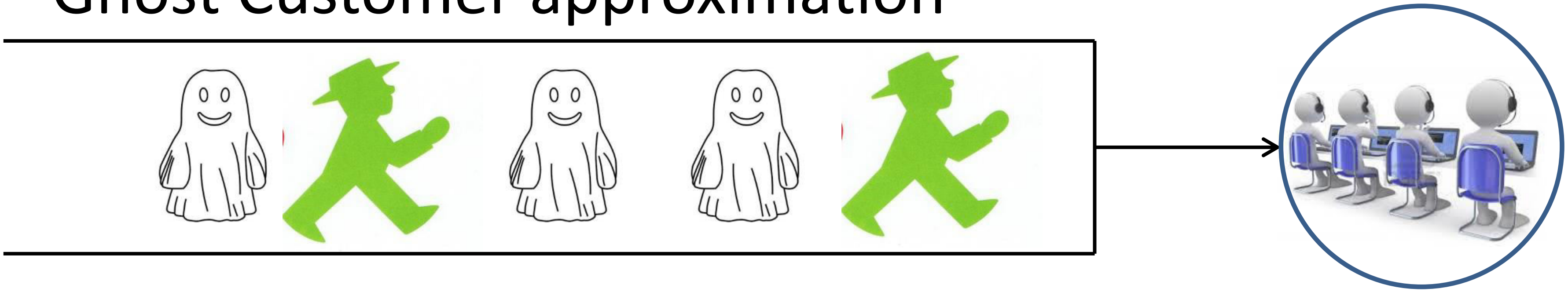
- Waiting Time vs Number in queue
- Example: The V model with FCFS
- Solution: The Erlang Approximation for the waiting time of the First in Line (FIL) (Koole et Al. (2012))
- M/M/1



$$p_{i,i-h} = \begin{cases} 1 - \sum_{h=0}^{i-1} \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h \; for\; i = h \\ \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h, \; for\; 0 \leq h < i. \end{cases}$$

With Abandonment ?

- The method is not directly applicable with abandonment
- Exception: The deterministic abandonment
- With $\frac{n}{\gamma}$ =τ, we reach the performance measures of the M/M/s+D queue.
- Extension to control in service problems
- Problem with other types of abandonment: a customer can abandon with a lower waiting time than the FIL.
- Ghost Customer approximation



Adaptation of the transition probabilities

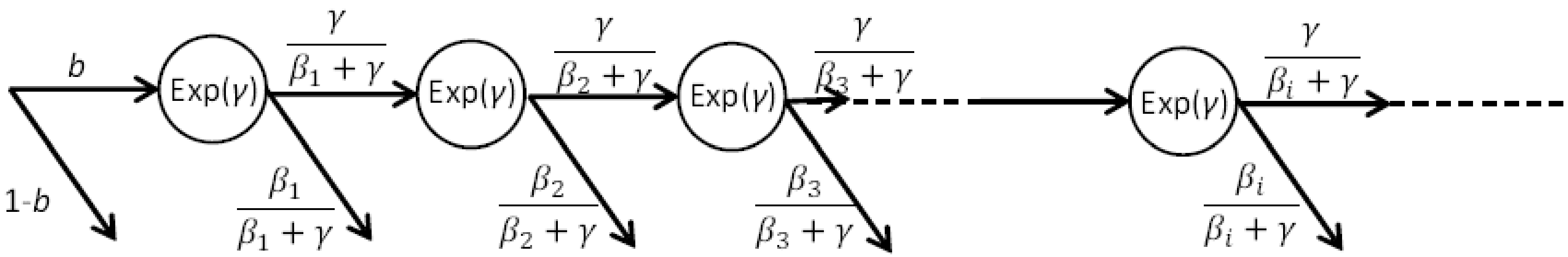
Recall that

$$p_{i,i-h} = \begin{cases} 1 - \sum_{h=0}^{i-1} \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h \; for\ i = h \\ \left(\frac{\lambda}{\lambda+\gamma}\right)\left(\frac{\gamma}{\lambda+\gamma}\right)^h, \; for\ 0 \le h < i. \end{cases}$$

Difficulties:

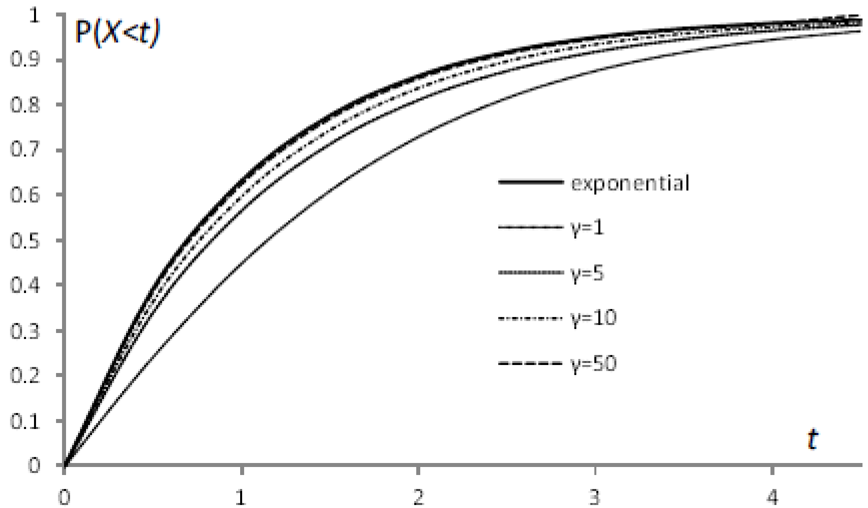
1. The state *i-h* does not only depend on the arrival of the following customer
2. The abandonment behavior is usually not exponential

General abandonment approximated by the Coxian distribution

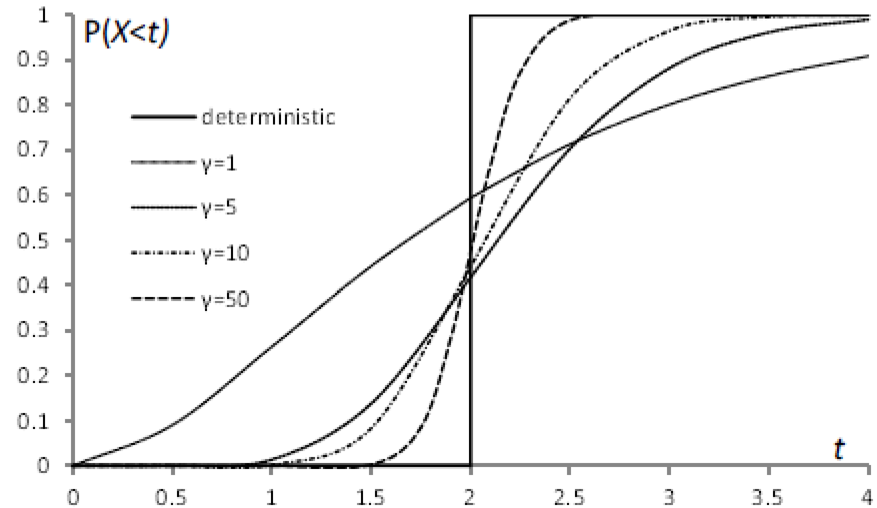


Convergence of our particular Coxian distribution

Theorem 1 *The random variable X_γ converges in distribution to any positive-valued distribution as γ tends to infinity.*



(a) Exponential Distribution with rate 1



(b) Deterministic distribition with parameter 2

Proposition 1 *With $b = 1$ and $\beta_i = \beta$ for $i > 0$, X_γ does not converge in probability to the exponential distribution with parameter β.*

Adaptation of the transition probabilities

After some computations …

$$p_{i,i-h} = \begin{cases} \prod_{k=1}^{i} q_k \ for\ i = h, i > 0 \\ (1 - q_{i-h}) \prod_{k=i-h+1}^{i} q_k, \ for\ 0 \leq h < i, \end{cases}$$

$$q_k = \frac{1}{1 + \frac{b\lambda}{\gamma} \prod_{j=1}^{k} \frac{\gamma}{\beta_j + \gamma}}, \ for\ k > 0.$$

- Another Numerical Method for the M/M/s+GI queue and more complex systems.

- Solution for Routing and Staffing problems with decisions based on the waiting time in systems with abandonment.

- No necessity for bounding the total event rate in the value functions.

$$\lambda + \gamma + s\mu + \max(\beta_k) = 1$$

Illustration: The V-Design

- Two FCFS queues, server have to choose which queue to prioritize, Hyperexponential abandonment

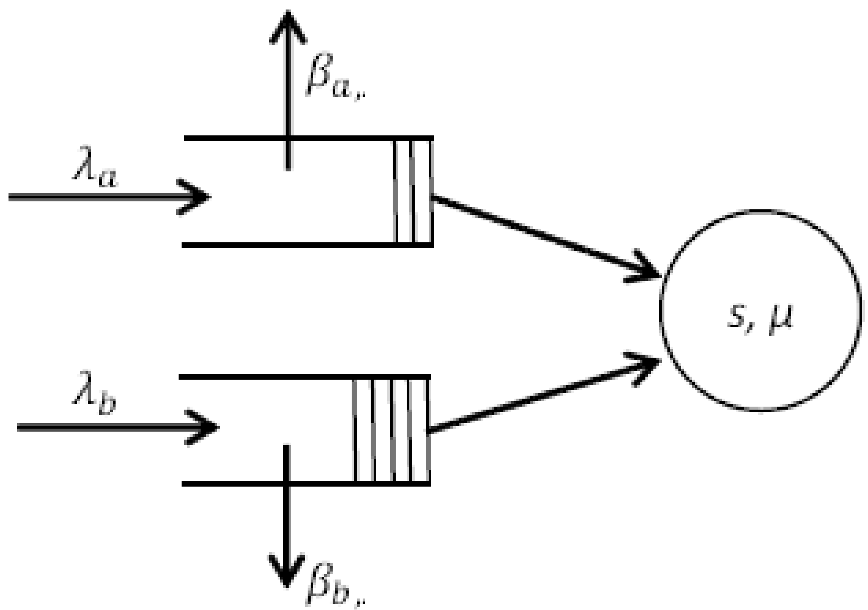
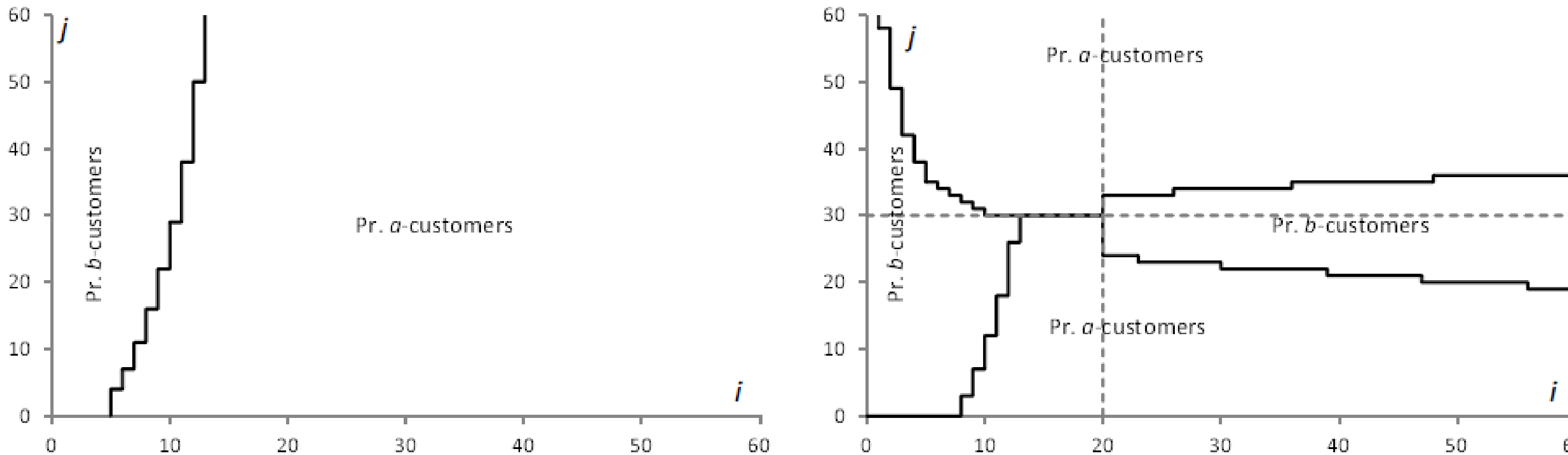


Illustration: The V-Design



(a) $c_a(i) = 5i/\gamma$, $c_b(j) = j/\gamma$, $i, j > 0$

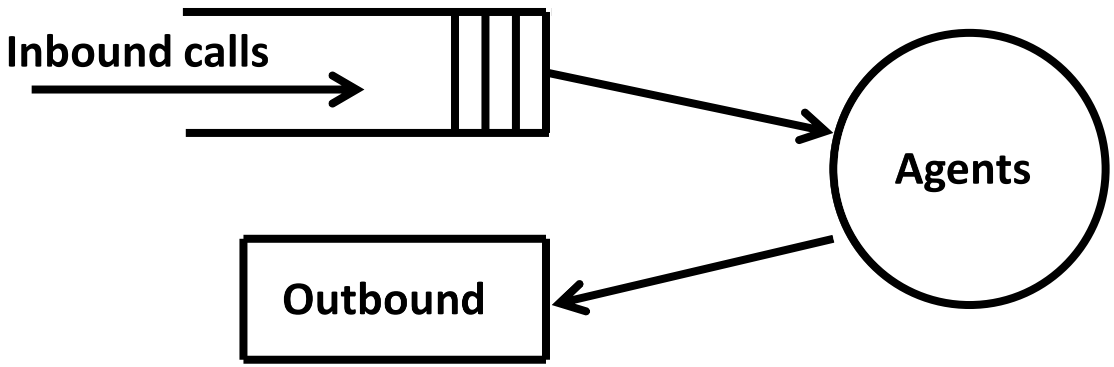
(b) $c_a(i) = 5$ for $i > 20$ and 0 otherwise, $c_b(j) = 1$, for $j > 30$ and 0 otherwise

Figure 4: Optimal policies ($\lambda_a = \lambda_b = 5$, $\mu = 1$, $s = 11$, $u_a = 10\%$, $\alpha_{a,1} = 1$, $\alpha_{a,2} = 5$, $u_b = 30\%$, $\alpha_{b,1} = 2$, $\alpha_{b,2} = 3$, $\gamma = 30$, $D = 120$)

The call center Blending Problem

Multichannel Call centers -> Combinations of urgent and non-urgent tasks

Time-dependent arrival rate, non-stationary analysis



Time sharing between inbound and outbound tasks ?

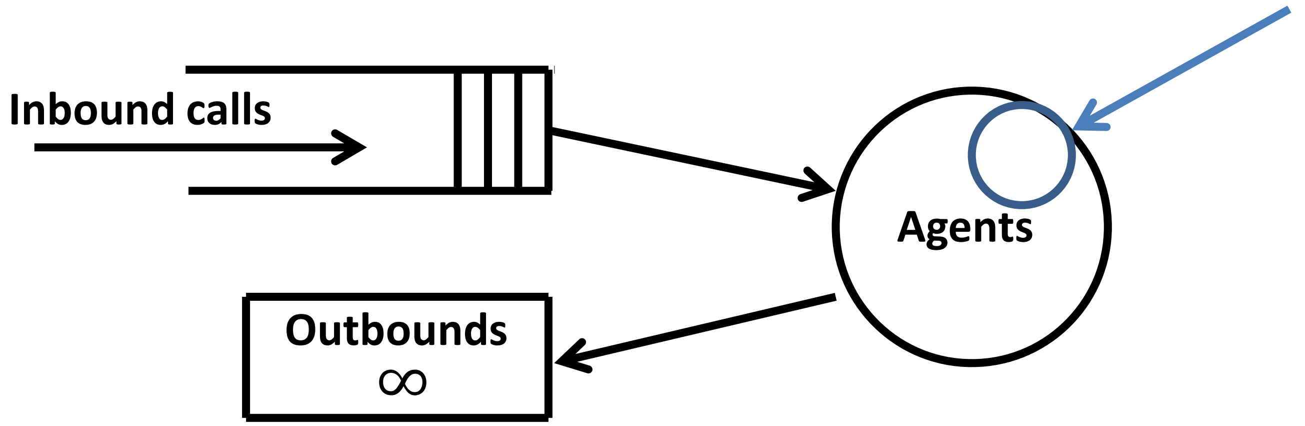
Propose « clever » Routing strategies ? (or non clever but cheap)

Motivation

Infinite amount of Outbound Tasks => Possibility of 100% utilization

=> Bad performance for Inbound Calls

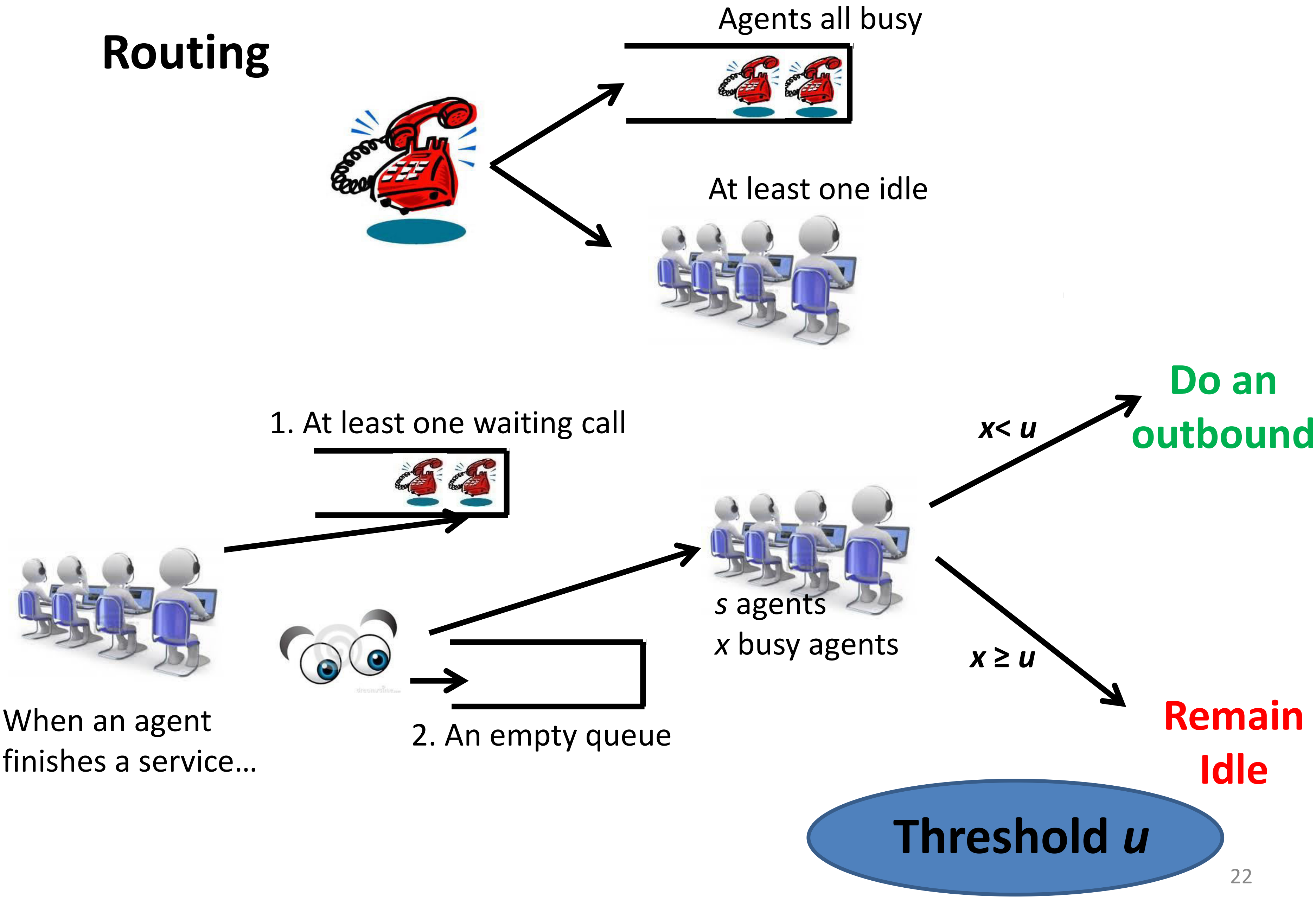
Solution: Reservation Threshold



Difficulties:

1. The reservation threshold is an integer
2. The arrival rate is time-dependent
3. The service level constraint on inbound calls has to be met for the whole working day

Adaptive Threshold Policies



Proof in the case of equal service rates, no abandonment

Bhulai, S. and Koole, G. (2003). A Queueing Model for Call Blending in Call Centers. *IEEE Transactions on Automatic Control*, 48:1434–1438.

Gans, N. and Zhou, Y.-P. (2003). A call-routing problem with service-level constraints. *Operations Research*, 51:255–271.

Performance comparison

Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., Ingolfsson, A., and Avramidis, A. (2007). Markov chain models of a telephone call center with call blending. *Computers & Operations Research*, 34(6):1616–1645.

Adaptive threshold

Bhulai, S., Farenhorst-Yuan, T., Heidergott, B., and Van der Laan, D. (2012). Optimal balanced control for call centers. *Annals of Operations Research*, 201(1):39–62.

Legros, B., Jouini, O., and Koole, G. (2015). Adaptive threshold policies for multi-channel call centers. *IIE Transactions*, 47:414–430.

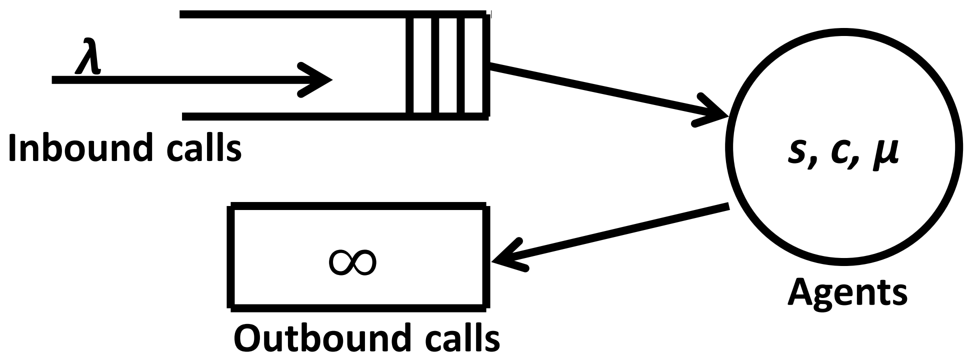
Staffing

Pang, G. and Perry, O. (2014). A logarithmic safety staffing rule for contact centers with call blending. *Management Science*, 61(1):73–91.

Main contribution

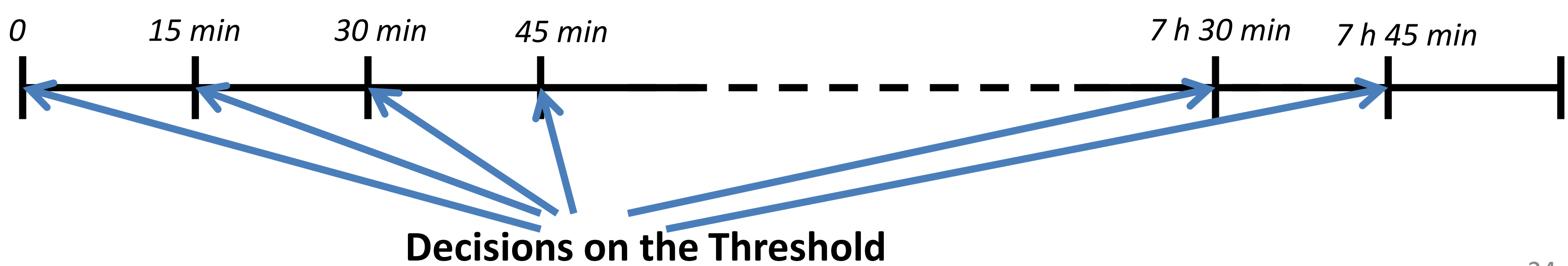
Efficient adaptive threshold policy easily implementable in the ACD

Modeling



- **<u>Throughput</u>** of the outbound tasks (*T*) ,
- Waiting time distribution of the calls (**<u>P(*W*<*t*)</u>**)

$$\begin{cases} \text{Maximize } T \\ \text{subject to } SL \geq \alpha, \end{cases}$$



Outline

- Stationary case

- Adaptive Threshold Policy (ATP)

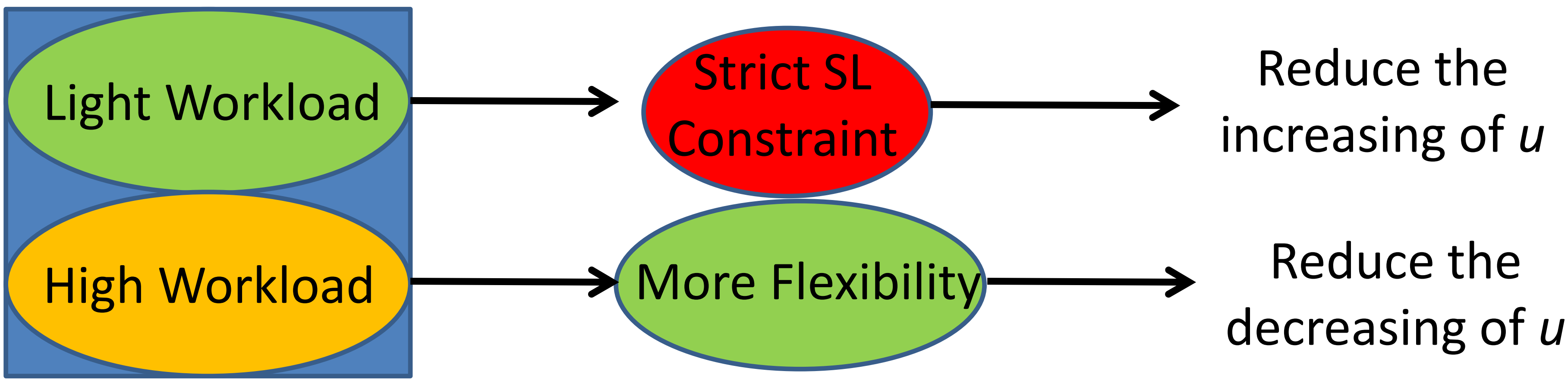
- Non stationary case

Performance measures

$$T(s,u,a) = \mu \left(\sum_{k=0}^{s-u} \frac{a^k u!}{(u+k)!} + \frac{a^{s-u} u!}{s!} \frac{a}{s-a} \right)^{-1} \left(u + \sum_{k=1}^{s-u} \frac{a^k u!}{(u+k-1)!} + \frac{a^{s-u+1} u!}{(s-1)!(s-a)} \right) - \lambda,$$

$$P(W < \tau) = 1 - C(s,u,a) e^{-\tau(s\mu - \lambda)}, \qquad C(s,u,a) = \frac{a^{s-u} u!}{s!(1-a/s)} \left(\sum_{k=0}^{s-u} \frac{a^k u!}{(u+k)!} + \frac{a^{s-u} u!}{s!} \frac{a}{s-a} \right)^{-1}.$$

Optimal solution on two Intervals

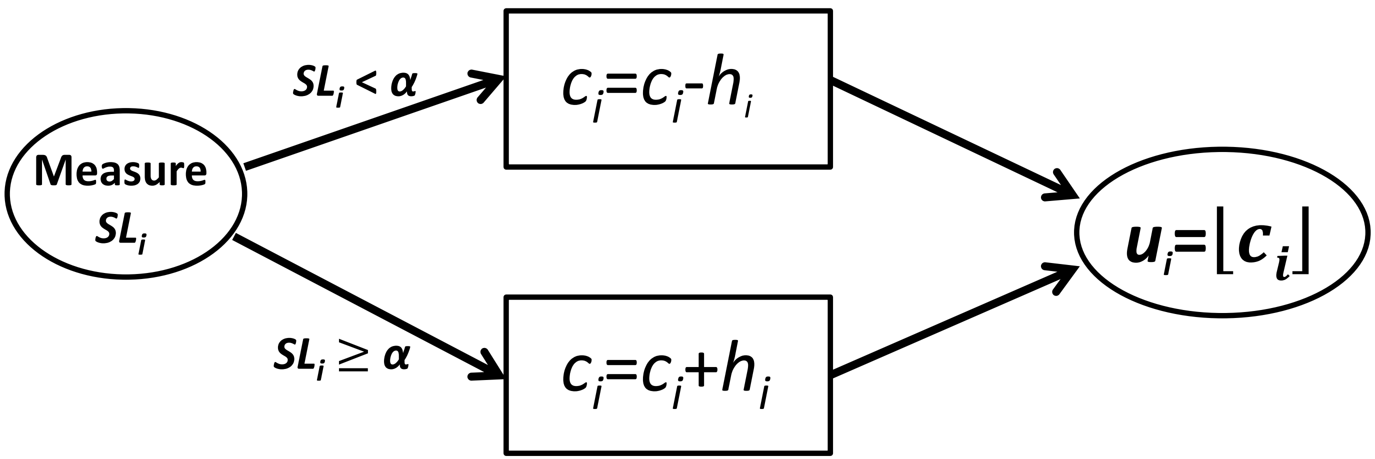


ATP Policy

c_i: the real threshold after i intervals

SL_i: the service level on calls over the last first i intervals

h_i: the value of change for c_i



Suggestion

$h_i = 1 - c_i/s$ if $SL_i \geq \alpha$ Slow increasing with high thresholds

$h_i = c_i/s$ if $SL_i < \alpha$ Slow decreasing with low thresholds

Comparison with Optimality

Theorem

Consider $0 \leq u_1, u_2 \leq s$ *such that* $SL(u_1) \leq \alpha \leq SL(u_2)$. *If it exists* $\gamma \in \mathbb{R}$ *for which* $(u_1, u_2) \in \arg\max_u T(u) + \gamma SL(u)$, *then randomizing between* u_1 *and* u_2 *is optimal.*

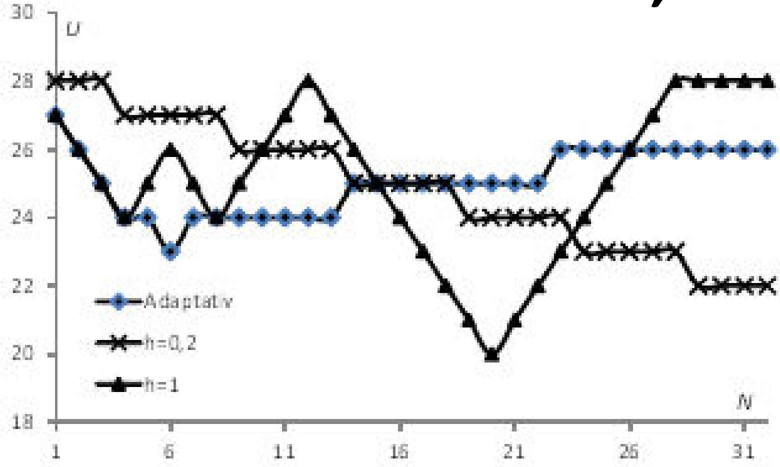
	Optimal c	Optimal T	ATP T	Difference
Scenario 1 ($\lambda = 4$, $\mu = \mu_0 = 0.2$, $s = 28$)	25.49	1.39	1.37	1.46%
Scenario 2 ($\lambda = 0.02$, $\mu = \mu_0 = 0.2$, $s = 1$)	0.13	0.02	0.02	0.00%
Scenario 3 ($\lambda = 18$, $\mu = \mu_0 = 0.2$, $s = 100$)	93.91	1.65	1.58	4.43%
Scenario 4 ($\lambda = 4$, $\mu = 0.27$, $\mu_0 = 0.15$, $s = 28$)	26.63	1.89	1.89	0.00%
Scenario 5 ($\lambda = 4$, $\mu = 0.17$, $\mu_0 = 1$, $s = 28$)	23.21	2.00	1.79	11.73%

Non Stationary Analysis

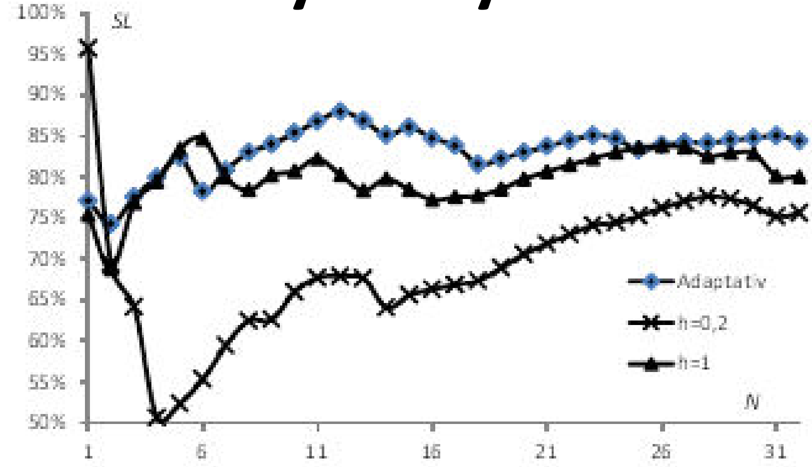
- Measure of the errors : $\overline{r_n} = \frac{\sum_{k=1}^{n} \mathrm{Max}(\alpha - \overline{SL_k}, 0)}{n}$
- Confidence interval for the proportion.
- Coefficient of aversion to the risk: *A*
- Function of comparison (Utility): $\overline{T_n} - A \times \overline{r_n},$

	h	$\overline{T}$	$\overline{SL}$	$\overline{r}$	U		h	$\overline{T}$	$\overline{SL}$	$\overline{r}$	U
	0.1	1.17	80.6%	0.0046	0.71		0.1	1.53	72.15%	0.0782	-6.3
	0.2	1.12	80.5%	0.0036	0.77		0.2	1.38	78.7%	0.0201	-0.63
Sc 1	0.5	1.04	80.1%	0.0032	0.72	Sc 2	0.5	1.23	81.4%	0.0063	0.60
	1	0.98	80.0%	0.0035	0.63		1	1.19	80.7%	0.0062	0.57
	ATP	1.09	80.7%	0.0027	0.82		ad	1.12	85.6%	0.0008	1.04

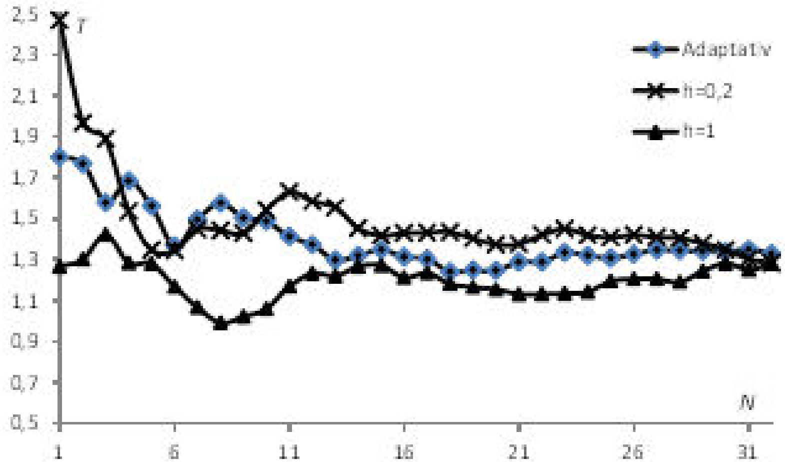
Numerical Illustration, non-stationary analysis



(a) Threshold



(b) Service Level



(c) Email Throughput

Conclusion

- Efficient adaptive threshold policy
- Comparison with the optimal policy with a constant stationary arrival rate
- Comparison with other intuitive policies under a non stationary analysis