Optimal control for resource allocation in a discrete queuing system.

Marc PIERRE, Richard COMBES, Salah EDDINE EL-AYOUBI
Laboratory of **S**ignals and **S**ystems, CentraleSupelec, Université Paris-Saclay

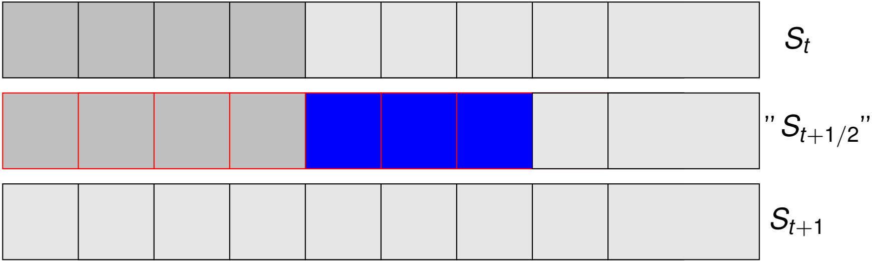
Introduction

- Our objective is to model a special case of queuing, where resource allocation is instantaneous, as is processing time, in order to optimize the cost of resource allocation.
- Discrete time queue, $(S_t)_{\mathbb{N}}$ represent the evolution over time of the number of customers in the queue, $(a_t)_{\mathbb{N}}$ the number of customers served at time t, and $(c_t)_{\mathbb{N}}$ the number of customers arriving at time t which are assumed *i.i.d*.

Introduction

This gives us the following model, $\forall t \in \mathbb{N}^*$:

$$S_{t+1} = (S_t + c_t - a_t)^+.$$



Markovian Decision Processes: a short introduction

- Using a Markovian Decision Processes (MDP) for our model.
- MDP are Markov chain with actions that influence the law of the chain.
- S the state space of the chain, A the space of possible actions.
- Being in a state s performing action a gives a cost $c(s, a)$ and change the state according to the law $P(.|s, a)$.

Markovian Decision Processes: a short introduction

- Find a decision rule, i.e. a sequence of functions d_t depending on the chain's trajectory up to time t, which minimize a cumulative cost.
- For a fixed $\gamma \in]0,1[$, we want to minimize :

$$v(s) := \mathbb{E}\Big[\sum_{t=0}^{+\infty} \gamma^t c(S_t, d_t(S_t)) | S_0 = s\Big]$$

$$Q(s,a) := \mathbb{E}\Big[\sum_{t=0}^{+\infty} \gamma^t c(S_t, d_t(S_t)) | S_0 = s, d_0(s) = a\Big].$$

Markovian Decision Processes: a short introduction

Markovian Decision Processes theory (see [Put94]) establishes the following theorem:

Theorem

There exists an application π^* from S in A, such that one of the minimizer $(d_t^*)_{\mathbb{N}}$ is of the form:

$$\forall t \in \mathbb{N} \; d_t = \pi^*.$$

This result greatly simplifies the search for an optimal decision rule, as this optimum is ultimately reached by a stationary, Markovian and non-random rule.

Model

In our case, the state space of the chain is described by $\mathbb{N}$, the action space by a set of the form $[\![0, a_{max}]\!]$, with $a_{max} \in \mathbb{N} \cup \{+\infty\}$, and the transition matrix P is fully described by the relation established previously:

$$S_{t+1} = (S_t + c_t - a_t)^+.$$

All that remains is to define the cost function. The idea is to translate this into a queue size L not to be exceeded, and a linear cost μ for the actions:

$$\forall (s, a) \in S \times A,\ r(s, a) = (s - L)^+ + \mu * a.$$

Many of our results can be generalized to reward functions of the form :

$$\forall (s, a) \in S \times A,\ r(s, a) = f(s) + g(a)$$

for any non decreasing f and g with f convex and g linear.

Shape of optimal policy

Non decreasing is optimal

We have the following property on the optimal policy π^*:

$$\forall s, \quad \pi^*(s+1) \geq \pi^*(s).$$

Although this result is intuitive, it is not easy to show.

Property

We have the following property one the optimal policy π^*:

$$\forall s, \quad \pi^*(s+1) \in \left\{\pi^*(s) + 1, 0, a_{max}\right\}.$$

Reducing the space Θ

We also want to avoid policies that allocate resources too late, otherwise we end up stuck in penalizing states. We also to avoid policies that allocate too much resources in 0.

Proposition

For $s \geq L$, and under the condition that $\gamma < \mu$ we have :

$$Q(s, 1) > Q(s, 0),$$

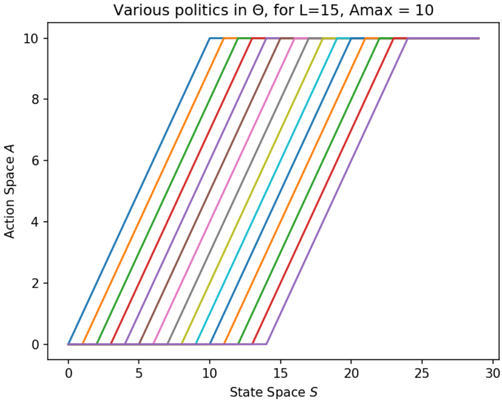
and in 0, we have the existence of $\tilde{a}$ such that:

$$\forall a < \tilde{a},\ Q(0, a) > Q(0, \tilde{a}).$$

The previous results, tells us that π^* is in space Θ where:

$$\Theta = \left\{\pi^\theta, \theta \in [\![-\tilde{a}, L[\![\ |\forall s \in \mathcal{S},\ \pi^\theta(s) = \min(a_{max}, (s-\theta)^+)\right\}.$$

Some Results



Some Results

In the settings of $a_{max} = \infty$, the queue can be stabilize, even with excessive arrivals, and Θ can be simplified to:

$$\Theta = \left\{\pi^{\theta}, \theta \in [\![-\tilde{a}, L[\![\;|\forall s \in \mathcal{S},\; \pi^{\theta}(s) = (s-\theta)^{+}\right\}.$$

We can compute the value of this politics by using the bellman equation, which leads to following property:

Value Computation

We have $\forall \pi^{\theta} \in \Theta$:

$$v^{\pi^{\theta}}(L) = f(L) + g(L-\theta) + \frac{\gamma}{1-\gamma}\mathbb{E}(h_{\theta}(Z))$$

where $h_{\theta}(Z) = f((\theta+z)^{+}) + g((\theta+z)^{+} - \theta)$.

Some Results

Laboratoire Signaux & Systèmes

Proposition

Given n i.i.d. samples $Z_1, \dots, Z_n$ from p_Z, define the estimator for $v^{\pi_\theta}(T)$ for $\theta \in \mathcal{L} := [\![-\tilde{a}, L[\![$

$$\hat{v}^{\pi_\theta}(L) = f(L) + g(L-\theta) + \frac{\gamma}{1-\gamma}\frac{1}{n}\sum_{i=1}^{n} h_L(Z_i)$$

the estimator for $v^\star(L)$

$$\hat{v}^\star(L) = \max_{\theta \in \mathcal{L}} \hat{v}^{\pi_\theta}(L)$$

and the corresponding estimate for $\pi^\star$

$$\hat{\pi}^\star = \pi^{\hat{\theta}} \text{ where } \hat{\theta} \in \arg\max_{\theta \in \mathcal{L}} \hat{v}^{\pi_\theta}(L)$$

Estimation Error

The procedure above yields a consistant estimate and its asymptotic error rate is upper bounded as

$$\limsup_{n\to\infty} \sqrt{n}\mathbb{E}\left(|\hat{v}^\star(u) - v^\star(u)|\right) \leq \sqrt{S^2 \ln(2|\mathcal{L}|)}$$

where σ_L^2 is the variance of $\frac{\gamma}{1-\gamma}h_L(Z)$, and $S^2 = \max_{L\in\mathcal{L}} \sigma_L^2$ is the largest variance.

Simulations

- Simulate a 5G base station servicing a set of users.
- User demand depends on a number of factors, including distance from the base station, antenna etc...
- Use our model to reduce allocation costs

Simulations

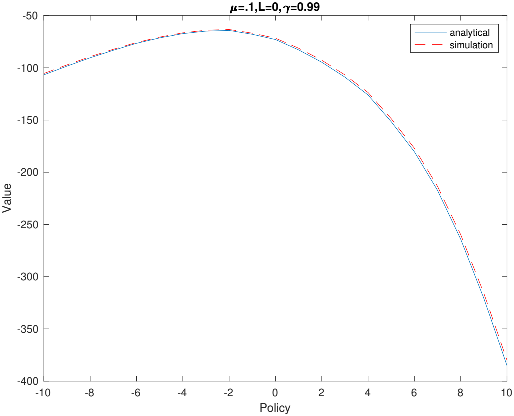


Figure: Value of the policy.

Simulations

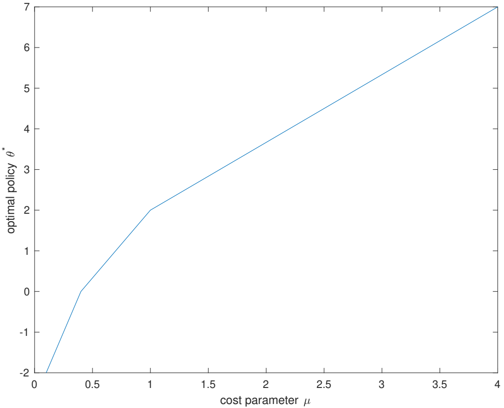


Figure: Optimal policy for different cost parameters μ..

Simulations

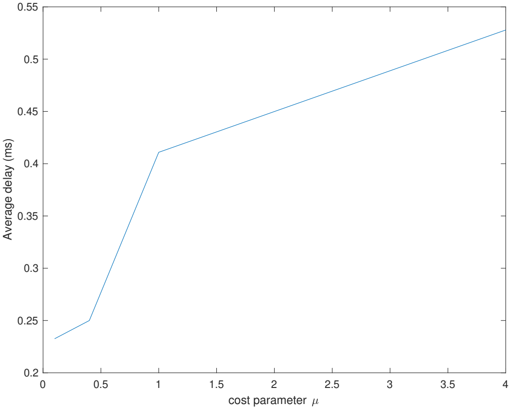


Figure: Impact of μ on the average delay

Conclusion

- Model our resource allocation problem as a reinforcement learning problem.
- One of the optimal policies is constant to 0 then increasing linearly.
- If we further assume that there is no limit to the number of resources we can allocate, these linearly increasing policies have a value that we can compute analytically.
- Estimating the value empirically with the analytical formula is not difficult and converges faster than trying to estimate the value function by the discounted sum.
- Computing an optimal policy can be done by minimizing a convex function.

Refferences

[Put94] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1st. USA: John Wiley & Sons, Inc., 1994. ISBN: 0471619779.

Thank you!

Thank you for your attention!