Asymptotic Optimality of the Binomial-Exhaustive Policy for Polling Systems with Large Switchover Times

CENTER FOR FINANCIAL & BUSINESS ANALYTICS
Data Science Institute

Yue Hu[+], Jing Dong[+], Ohad Perry[*]
+ Columbia Business School
* Northwestern University

Data Science Institute
COLUMBIA UNIVERSITY

The Model and Problem Statement

A **polling system** is a stochastic queueing network where several queues are served by one switching server

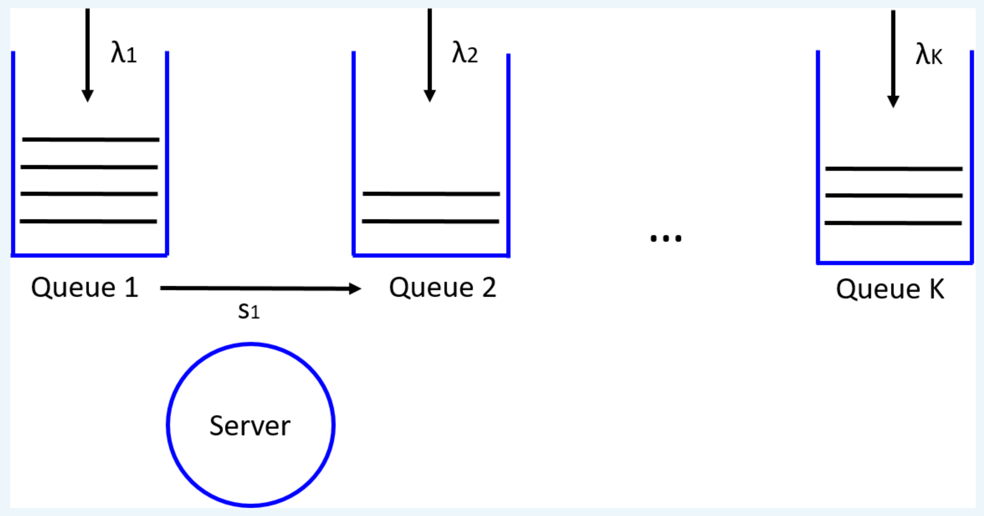

Figure 1. Polling system with K queues

- A single server switches between $K \geq 2$ queues indexed by $1 \leq i \leq K$
- The server attends to the queues in a periodic cycle of a stages defined by the polling table $p: \{1,\ldots,a\} \rightarrow \{1,\ldots,K\}$
  - $p(j) = i$ – the server visits queue i at stage j within a cycle
  - $a \geq K$ – a queue can be visited multiple times in a cycle
  - E.x. server routing {1, 3, 2, 3, 2 | 1, 3, 2, 3, 2 |…}, K = 3, a = 5
  - **Fix the polling table, and optimize the service (server-switching) policy**
- Poisson arrival process of jobs with rate λ_i to queue i
- IID service times V_i with mean $1/\mu_i$ at queue i
- Necessary stability condition: $\sum_{i=1}^{K} \lambda_i / \mu_i < 1$
- Large switchover times S_j is incurred when server switches from stage j

Stochastic dynamics:

- $Q_i(t)$ – number in queue i at time t
- $Z(t)$ – the location of the server at time t
  - $Z(t) = j$ – server is at stage j
  - $Z(t) = \ominus_j$ – server is switching from stage j to stage j+1 (modulo a)
- The stochastic dynamics are characterized via the process

$$X(t) := (Q_i(t), Z(t), 1 \leq i \leq K), t \geq 0$$

The problem:

- Our goal is to find a service policy that minimizes the long-run average holding cost among all admissible controls that are non-idling, non-anticipative, and discrete-Markov (given the queue length at the polling instant, the service policies do not depend on the history of the system-level process up to that time instant)

The Fluid Approximation

- Consider a hybrid dynamical system (HDS) characterized by $x(t) := (q(t), z(t))$, where x(t) is a hybrid of the piece-wise linear continuous number-in-system process q(t), and the discrete server-location process z(t), characterized via

$$\dot{q}_i(t) = \lambda_i - \mu_i 1\{p(z(t) = i, q_i(t) > 0\}$$

- The fluid control problem aims to find a control that minimizes

$$c := \limsup_{t \to \infty} \frac{1}{t} \int_0^t \varphi(q(s))ds,$$

where φ is a general cost rate function that is non-decreasing and continuous

The Fluid Control Problem

Fluid periodic equilibrium:

- Any periodic equilibrium can be mapped to a vector of service proportions (r_j, $1 \leq i \leq a$)

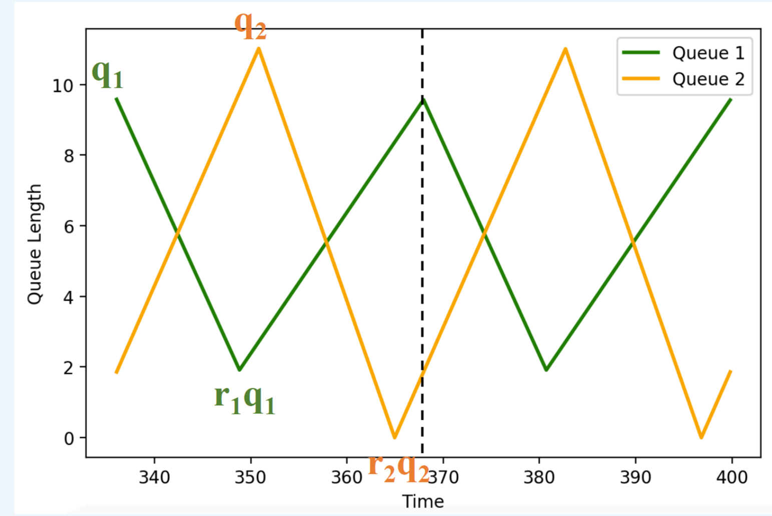

Figure 2. Example periodic equilibrium with two queues and server routing {1, 2 | 1, 2 |…}

- **Stage-Based Proportion Reduction Control (SB-PR):**
  - For fixed ($r_j \in [0, 1]$, $1 \leq j \leq a$), if $p(j) = i$, then reduce q_i (the value of queue i at the beginning of service) to $r_j q_i$
  - After serving state j, switch to stage j + 1 in the polling table

Theorem: Independently of initialization,
- the HDS converges to a unique periodic equilibrium under SB-PR with any non-trivial control parameters
- the HDS converges to the desired periodic equilibrium under the corresponding SB-PR

- It follows that the SB-PR control is optimal for the fluid control problem

Asymptotic Optimality

Tranlsate the fluid control for the stochastic system:

- **Binomial-Exhaustive Policy:**
  - At the polling epoch of stage j, for each customer present in queue p(j), serve that customer (together with all the new arrivals during his/her service time) with probability $1 - r_j$
  - After serving stage j, switch to stage j+1 in the polling table
  - On average, reduce queue p(j) from $Q_{p(j)}$ to $r_j Q_{p(j)}$

Large switchover time scaling:

- Consider a sequence of systems indexed by n
  - Large switchover times: $\bar{S}_j^n \Rightarrow s_j$, $E[\bar{S}_j^n] \rightarrow s_j$
  - Temporal and special scaling: $\bar{X}^n(t) := (Q(nt)/n, Z(nt))$
  - Scaled long-run average cost: $\bar{C}^n := \lim_{t \to \infty} \frac{1}{t} \int_0^t \varphi(\bar{Q}^n(t))ds$
- Let π_*^n denote the binomial-exhaustive policy translated from the optimal fluid control for the n^{th} stochastic system, and c_* denote the optimal fluid cost

Theorem: Under regularity conditions on the service and switchover time distributions,
- under any sequence of admissible controls π^n, $\liminf_{n \to \infty} \bar{C}^n \geq c_*$ w.p.1
- under the sequence of fluid-translated binomial-exhaustive policies π_*^n, $\bar{C}^n \Rightarrow c_*$ w.p.1

- It follows that the binomial-exhaustive policy is asymptotically optimal