Learning Optimal Queue Dispatch Policies

Fethi Bencherki * **Anders Rantzer** *

* *Department of Automatic Control, Lund University, Box 118, 22100, Lund, Sweden (e-mail: fethi.bencherki@control.lth.se).*

Abstract: This work presents a dual controller that learns the optimal server in a multi-server queueing system under process disturbances. We formulate an optimization problem with linear cost, linear dynamics, and an equality constraint on the dispatcher policy. A model-free, data-driven equation is constructed to enable both policy evaluation and update.

Keywords: Queueing systems; Dual control.

1. INTRODUCTION

Queueing systems find wide applications in communication networks, supply chains, and resource allocation. Optimal control of such systems is crucial for improving efficiency and reducing operational costs (Bertsimas and Kim, 2023).

The optimal control of queueing systems with unknown service rates has gained attention due to its relevance in applications like cloud computing, where service rates are variable and uncertain due to interruptions and slowdowns (Choudhury et al., 2021). While traditional approaches assume known parameters (Tassiulas and Ephremides, 1990), recent efforts have explored learning-based control to address this challenge (Liang and Modiano, 2018; Krishnasamy et al., 2021; Stahlbuhk et al., 2021). In this work, we adopt a dual control framework to develop an adaptive algorithm for optimal job dispatching under such uncertainty and disturbances

2. PROBLEM SETUP

2.1 A multi-unit processing network model

We consider a multi-unit processing network model consisting of n queueing units, with jobs awaiting to be processed at each unit. The number of external jobs arriving at the network is a deterministic quantity denoted by λ, to be forwarded by the dispatcher to the different units for processing. We let $[u_t]_i$ denotes the portion of jobs routed to unit i. The units process the jobs at different processing rates denoted by $\eta_i > 1$ for unit $i \in \{1, \dots, n\}$. This setup is illustrated in Figure 1. We let the vector $[x_t]_i$ represent the number of jobs awaiting at unit i at time t, and adhering to the following linear dynamics

$$[x_{t+1}]_i = \tfrac{1}{\eta_i}[x_t]_i + [u_t]_i, \quad \eta_i > 1, \tag{1}$$

and in aggregation across all units

$$x_{t+1} = Mx_t + u_t, \quad M \coloneqq \operatorname{diag}\left(\eta_1^{-1}, \dots, \eta_n^{-1}\right). \tag{2}$$

Note that according to (1), unit i possessing a large processing rate η_i indicates that it enjoys a faster linear convergence with a rate of $0 < \eta_i^{-1} < 1$. The inputs, on the other hand, adhere to the constraints

$$\mathbf{1}^\top u_t = \lambda \quad \text{and} \quad u_t \geq 0 \quad \forall t. \tag{3}$$

The equality constraint on the inputs could be justified in the sense that we would like the incoming jobs λ at time t to be fully distributed for processing. The inequality constraint on the other hand to signify that a job is a non-negative quantity.

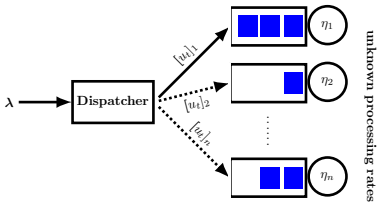

Fig. 1. An illustration of the multi-unit processing network.

2.2 An optimal control problem

A typical performance objective is to minimize the sum of awaiting jobs across all units. This gives rise to the following optimization problem with a discount factor $\gamma \in (0, 1)$

$$\begin{aligned} \text{Minimize} \quad & \textstyle\sum_{t=0}^{\infty} \gamma^t \mathbf{1}^\top x_t \text{ over } \{u_t\}_{t=0}^{\infty} \\ \text{subject to} \quad & x_{t+1} = Mx_t + u_t, \\ & \mathbf{1}^\top u_t = \lambda \quad \text{and} \quad u_t \geq 0, \quad \forall t. \end{aligned} \tag{4}$$

2.3 Optimal policy for known processing rates

When the processing rates are known, problem (4) can be solved using dynamic programming (Bellman, 1966), leading to the Bellman equation

$$J(x) = \min_{u \in \mathcal{U}(\lambda)} \left[\mathbf{1}^\top x + \gamma J(Mx + u)\right], \tag{5}$$

where $\mathcal{U}(\lambda)$ denotes the set of admissible inputs. Assuming an affine form for the value function, $J(x) = p^\top x + p^0$, with $p \in \mathbb{R}^n_+$, we obtain the solution

$$p^0 = \frac{\gamma\lambda \min_i(p_i)}{1-\gamma}, \quad p_i = \frac{\eta_i}{\eta_i - \gamma}. \tag{6}$$

The optimal policy is $u^* = \lambda e_i$, where $i = \arg\max_j \eta_j$, indicating that jobs should be routed to the fastest server.

⋆ This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation.

2.4 Model-free optimal control via Q-factor

The cost-to-go from time t under policy u, starting from state x_t, is defined as $J(x_t) = \min_u \sum_{k=t}^{\infty} \gamma^{k-t} \mathbf{1}^\top x_k$. The optimal Q-function (Bradtke et al., 1994) is $Q(x_t, u_t) = \mathbf{1}^\top x_t + \gamma J(x_{t+1})$, and assuming the value function is affine, this yields $Q(x_t, u_t) = \mathbf{1}^\top x_t + \gamma p^\top (M x_t + u_t) + \gamma p^0$. This can be written compactly as $Q(x_t, u_t) = q^\top \begin{bmatrix} x_t \\ u_t \end{bmatrix} + q^0$, where $q = \begin{bmatrix} q^x \\ q^u \end{bmatrix} = \begin{bmatrix} \mathbf{1} + \gamma M p \\ \gamma p \end{bmatrix}$, and $q^0 = \gamma p^0$. Notably, $q^u = \gamma p = \gamma q^x$, and $q^0 = \frac{\gamma^2 \lambda}{1-\gamma} \min_i [q^x]_i$. The Q-function also satisfies the Bellman equation in Q-factor form (Sutton and Barto, 2018), given by

$$Q(x_t, u_t) = c(x_t, u_t) + \gamma \min_{u_{t+1} \in \mathcal{U}(\lambda)} Q(x_{t+1}, u_{t+1}), \quad (7)$$

with the optimal policy as $u_t^* = \arg\min_{u_t \in \mathcal{U}(\lambda)} Q(x_t, u_t)$. Substituting the affine form of the Q-function into (7) and simplifying, we obtain

$$\left(q - \begin{bmatrix} \mathbf{1} \\ 0 \end{bmatrix} \right)^\top \begin{bmatrix} x_t \\ u_t \end{bmatrix} = \gamma q^\top \begin{bmatrix} x_{t+1} \\ 0 \end{bmatrix} + \beta, \quad (8)$$

where $\beta \coloneqq (\gamma - 1) q^0 + \gamma \lambda \min_i [q^u]_i$. Using the expressions for q^0 and q^u, we find that $\beta = 0$. Substituting $q^u = \gamma q^x$ into (8) with $\beta = 0$, we arrive at the model-free equation

$$(q^x - \mathbf{1})^\top x_t = \gamma (q^x)^\top (x_{t+1} - u_t). \quad (9)$$

Collecting data over time, we stack t such relations to obtain

$$(q^x - \mathbf{1})^\top [x_0 \ \cdots \ x_{t-1}] = \gamma (q^x)^\top [x_1 - u_0 \ \cdots \ x_t - u_{t-1}]. \quad (10)$$

Multiplying (10) on the right by $Z_t^\top$, where $Z_t \coloneqq [x_0 \ \cdots \ x_{t-1}]$, and defining the empirical data correlation matrices

$$\Sigma_t = \sum_{k=0}^{t-1} x_k x_k^\top + \Sigma_0, \quad \bar{\Sigma}_t = \sum_{k=0}^{t-1} (x_{k+1} - u_k) x_k^\top, \quad (11)$$

with regularization term $\Sigma_0 \succ 0$, yields the linear data-driven equation

$$(q^x - \mathbf{1})^\top \Sigma_t = \gamma (q^x)^\top \bar{\Sigma}_t, \quad (12)$$

from which we can compute an estimate of the optimal parameter q^x directly from data using

$$\begin{cases} q_t^x \coloneqq (I - \gamma \Sigma_t^{-1} \bar{\Sigma}_t^\top)^{-1} \mathbf{1}, \\ q_t^u \coloneqq \gamma q_t^x, \\ q_t^0 \coloneqq \dfrac{\gamma^2 \lambda}{1-\gamma} \min\limits_i [q_t^x]_i, \end{cases} \quad (13)$$

where q_t^x denotes the data-driven estimate of q^x, forming the basis of the proposed policy construction as will be discussed next.

2.5 Problem Formulation

In this letter, we assume that the dispatcher does not know the processing rates and has to learn them while scheduling dispatching policies in an online fashion. After dispatching a job at time t, it receives a noisy observation of the number of awaiting jobs across all units at the end of time slot $[t, t+1]$, denoted by x_{t+1} and given by

$$x_{t+1} = M x_t + u_t + w_t. \quad (14)$$

Inspired by (13), we propose and analyze the performance of the following dispatching policies in controlling system (14)

$$\begin{cases} \Sigma_t = \Sigma_{t-1} + x_{t-1} x_{t-1}^\top, \quad \Sigma_0 \succ 0 \\ \bar{\Sigma}_t = \bar{\Sigma}_{t-1} + (x_t - u_{t-1}) x_{t-1}^\top, \quad \bar{\Sigma}_0 = 0 \\ q_t^x = (I - \gamma \Sigma_t^{-1} \bar{\Sigma}_t^\top)^{-1} \mathbf{1}, \\ u_t = \lambda e_{i_t} + \epsilon_t \quad \text{where} \quad i_t = \arg\min\limits_j [q_t^x]_j. \end{cases} \quad (15)$$

At each step, a control input u_{t-1} is applied at state x_{t-1}, resulting in a new state x_t, and the triplet (x_{t-1}, u_{t-1}, x_t) is used to update the correlation matrices $(\Sigma_t, \bar{\Sigma}_t)$ in policy (15). An estimate q_t^x is then computed, from which the first component of the next control u_t is derived. The second component, ϵ_t, ensures exploration and serves as a probing action in the spirit of dual control (Wittenmark, 1995). The linear model in (2) is robustified by additive disturbances w_t, which capture unmodeled dynamics, uncontrolled traffic, and on/off server behavior (Zhou and Doyle, 1998), motivating our avoidance of statistical assumptions on w_t.

REFERENCES

Bellman, R. (1966). Dynamic programming. *Science*, 153(3731), 34–37.

Bertsimas, D. and Kim, C.W. (2023). Optimal control of multiclass fluid queueing networks: A machine learning approach. *arXiv preprint arXiv:2307.12405*.

Bradtke, S.J., Ydstie, B.E., and Barto, A.G. (1994). Adaptive linear quadratic control using policy iteration. In *Proceedings of 1994 American Control Conference-ACC'94*, volume 3, 3475–3479. IEEE.

Choudhury, T., Joshi, G., Wang, W., and Shakkottai, S. (2021). Job dispatching policies for queueing systems with unknown service rates. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 181–190.

Krishnasamy, S., Sen, R., Johari, R., and Shakkottai, S. (2021). Learning unknown service rates in queues: A multiarmed bandit approach. *Operations research*, 69(1), 315–330.

Liang, Q. and Modiano, E. (2018). Minimizing queue length regret under adversarial network models. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(1), 1–32.

Stahlbuhk, T., Shrader, B., and Modiano, E. (2021). Learning algorithms for minimizing queue length regret. *IEEE Transactions on Information Theory*, 67(3), 1759–1781.

Sutton, R.S. and Barto, A.G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tassiulas, L. and Ephremides, A. (1990). Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *29th IEEE Conference on Decision and Control*, 2130–2132. IEEE.

Wittenmark, B. (1995). Adaptive dual control methods: An overview. *Adaptive Systems in Control and Signal Processing 1995*, 67–72.

Zhou, K. and Doyle, J.C. (1998). *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ.