Machine Learning-based Queuing Model Regression - Example Selection, Feature Engineering and the Role of Traffic Intensity

Roland Braune

Department of Business Decisions and Analytics
University of Vienna
roland.braune@univie.ac.at

1 Extended Abstract

The subject of this contribution is the performance prediction of queuing models based on advanced (multiple) regression techniques. The main motivation for this kind of approach is to instantly provide high-quality estimates for key performance indicators of complex queuing systems that could only be tackled by time-consuming simulation procedures otherwise. Examples include queues with general or degenerate arrival or service time distributions, non-homogeneous arrival processes, transient behavior and queuing networks with limited waiting space at nodes. Particularly when embedded in a simulation-based optimization framework, the run-time of such models becomes a crucial factor. Learning a regression model for the queuing system under consideration and then using it as a surrogate model for its simulation counterpart can therefore lead to considerable time savings during the run of a heuristic optimization algorithm.

Related work in this area is mainly based on generic approaches, like stochastic kriging (see, e.g., [5] and [2]), response surface methodology (RSM) [1], curve fitting techniques (e.g., [8] and [7]), and also Bayesian inference [6]. The incorporation of typical properties of queuing models into a regression approach is rarely addressed in the scientific literature. For example, Ouyang and Nelson [3] propose a logistic regression-based approach for the prediction of blocking probabilities, while Senderovich et al. [4] rely on non-linear regression and regression trees for delay prediction in service processes.

The goal of the contribution at hand is a regression analysis of various different types of queuing systems, including queuing networks that violate conditions of tractable models like Jackson networks. The regression models take as an input certain configuration parameters of the system, like arrival or service rates, or the number of servers at each node. Performance measures of interest include the average number of customers in the system, the mean waiting and sojourn time, and various quantities derived from the probability distribution of waiting (or sojourn) times, like percentiles. It has to be emphasized that the proposed approach is generic in the sense that no a priori assumptions with regard

to the mathematical structure of the regression itself have to be made. Queuing-specific aspects, on the other hand, are involved by defining an appropriate and effective set of custom features.

The first part of the study shows the validity of the approach, based on queuing models for which closed-form analytical solutions exist. It turns out that non-linear regression techniques, in particular support vector regression and kernel ridge regression are able to achieve almost perfect fits. The introduction of non-linear combinations of simple features like the arrival/service rates and the number of servers appears to be indispensable nevertheless. A traffic intensity-driven data generation scheme is compared to a purely random one. The experimental coverage is then extended towards queuing networks of different kinds. Finally, potential and immediate application scenarios in ongoing projects, such as the optimization of a flexible manufacturing system and a public transport network, are sketched.

References

[1] Paul D. Berger, Robert E. Maurer, and Giovana B. Celli. *Introduction to Response-Surface Methodology*, pages 533–584. Springer International Publishing, Cham, 2018.

[2] Jack P.C. Kleijnen. Regression and kriging metamodels with their experimental designs in simulation: A review. *European Journal of Operational Research*, 256(1):1 – 16, 2017.

[3] H. Ouyang and B. L. Nelson. Simulation-based predictive analytics for dynamic queueing systems. In *2017 Winter Simulation Conference (WSC)*, pages 1716–1727, 2017. logistic regression, prediction of probability that system is blocked after some time.

[4] Arik Senderovich, Matthias Weidlich, Avigdor Gal, and Avishai Mandelbaum. Queue mining for delay prediction in multi-class service processes. *Information Systems*, 53:278–295, 2015. regression approach for queueing delays, non-linear regression and regression trees.

[5] Haihui Shen, L. Jeff Hong, and Xiaowei Zhang. Enhancing stochastic kriging for queueing simulation with stylized models. *IISE Transactions*, 50(11):943–958, 2018.

[6] C. Sutton and M. I. Jordan. Bayesian inference for queueing networks and modeling of internet services. *The Annals of Applied Statistics*, 5(1):254–282, 2011.

[7] C. Yan, L. Mönch, and S. M. Meerkov. Characteristic curves and cycle time control of re-entrant lines. *IEEE Transactions on Semiconductor Manufacturing*, 32(2):140–153, May 2019.

[8] Feng Yang, Bruce Ankenman, and Barry L. Nelson. Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Research Logistics (NRL)*, 54(1):78–93, 2007.