Queueing Systems (2022) 100:337–339
https://doi.org/10.1007/s11134-022-09800-8

Many-server limits for service systems with dependent service and patience times

Pascal Moyal[1] · **Ohad Perry**[2]

Received: 6 February 2022 / Accepted: 28 February 2022 / Published online: 3 May 2022

1 Introduction We consider the open problem of establishing a functional weak law of large numbers (FWLLN) for the queue process, and a corresponding weak law of large numbers (WLLN) for the stationary distribution, of service systems in which the service and patience times of each customer are dependent random variables. In particular, the systems we have in mind are of the $GI/GI/n + GI$ type, having a renewal arrival process of statistically homogeneous customers that are served by n statistically homogeneous agents, in addition to customer abandonment from the queue (the $+GI$ in the notation). However, unlike the typical $GI/GI/n+GI$ system, we want to consider the system under the assumption that the service requirement of each customer depends on that customer's patience for waiting in the queue. Such systems are prohibitively hard to analyze even if the arrival process is Poisson, and the (marginal) distributions of the service and the patience times are exponentials, because the queue process does not admit a finite-dimensional Markov representation.

It is significant that the dependence between service and patience changes the queueing dynamics significantly, as can be deduced immediately from the approximation for the stationary distribution in (2) below. Indeed, in [13] it is proved that, if the two random variables are perfectly correlated and both are (marginally) exponentially distributed, and if the arrival process is Poisson, then the queue behaves asymptotically as if there is no abandonment at all under diffusion scaling.

Background When considering service systems, human behavior must be taken into account in order to properly analyze and optimize such systems. In particular, customer abandonment plays an important role in the modeling of service systems, because abandonment has fundamental impacts on the queueing dynamics. The typical approach to modeling service and patience times is to assume that each customer arrives to the system with a service requirement and patience, both being random

✉ Ohad Perry
ohad.perry@northwestern.edu

Pascal Moyal
pascal.moyal@univ-lorraine.fr

1 IECL, Université de Lorraine, Metz, France

2 Industrial Engineering and Management Science, Northwestern University, Evanston, USA

variables that are independent from all other random variables describing the system, and in particular from each other. However, human behavior is clearly more complex than this naive modeling approach. It stands to reason that in many practical settings, the service requirement of each customer depends on that customer's patience or on the delay she experiences in queue. Indeed, dependence between the service times and the delay in queue has been empirically observed in hospitals [2], restaurants [3], retail stores [1], and contact centers [8].

2 Problem statement As described above, we consider the $GI/G_{dep}/n + G_{dep}$ in which the service and patience times are dependent (hence the 'dep' in the subscripts). We denote by σ_k and D_k the service and patience times of the kth arrival after time 0, respectively, and assume that the sequence $\{(\sigma_k, D_k) : k \geq 1\}$ is i.i.d. in $\mathbb{R}^2_+$ with joint density f. We denote by λ^n the arrival rate to system n, and assume that $\lambda^n/n \to \lambda > 0$ as $n \to \infty$. We propose employing the measure-valued approach taken in [7] and in [5] which proved FWLLNs for the $GI/GI/n$ and the $GI/GI/n + GI$ models, respectively.

Specifically, for any $t \geq 0$, let W^n_t, S^n_t and X^n_t be the number of customers in queue, in service, and in the overall system (queue + service) at time t. For $i = 1, \ldots, W^n_t$ and $j = 1, \ldots, S^n_t$, let w^i_t be the time spent in line by the ith customer in queue, s^j_t be the time spent in service by the jth customer in service at t, and consider the two point measures

$$\eta^n_t = \sum_{i=1}^{W^n_t} \delta_{w^i_t} \quad \text{and} \quad \nu^n_t = \sum_{j=1}^{S^n_t} \delta_{s^j_t}. \tag{1}$$

Then, a FWLLN would state that, under appropriate regularity conditions, the fluid-scaled sequence $\{(\eta^n, \nu^n, X^n)/n : n \geq 1\}$ converges weakly to the unique solution of a deterministic integral equation. We believe that the *fluid limit* $\bar{X}$ of the latter sequence is equivalent to the two-parameter fluid model (derived directly without using asymptotic arguments) in [11]; see [10, Chapter 3].

It is also argued in [11] that the fluid model has a stationary point x^*, of the form

$$x^* = 1 + \lambda \int_0^w (1 - F_T(x))\, dx, \tag{2}$$

where F_T is the cdf of the D_k's, and w is the unique solution to

$$\lambda \int_w^\infty \int_0^\infty x f(x, y) dx dy = 1,$$

see also [9]. Then, letting $X^n(\infty)$ denote a random variable having the steady-state distribution of the process X^n, we would like to prove the interchange of limits, by showing that $X^n(\infty)/n$ converges weakly in $\mathbb{R}$ to x^*, whenever x^* is the unique stationary point.

3 Discussion It is significant that the FWLLN for the $G/GI/n + GI$ queue, proved in [5], relies heavily on the assumption that the service and patience times of each customer are independent. In particular, the martingale representation of the queueing

dynamics employed in [5] fails to hold if this is not the case; see [5, Proposition 5.1]. Thus, the FWLLN we want to prove does not follow from existing results, nor can the analysis in [5] be directly generalized to our setting. A possible approach is to consider an equivalent system for the $G/G_{dep}/n + G_{dep}$ in which the service time of each customer depends on that customer's waiting time in queue, and is independent of that customer's patience. The existence of such an equivalent system (in the sense that the queues in both systems have the same distribution) was recently proved in [12]. This representation is simpler in that it somewhat decouples the dynamics of the two processes η^n and ν^n, however, it would require to adapt the framework of [5], by keeping track of *residual* service times (as, e.g., in [4]), rather than ages.

Proving the WLLN for the stationary queue is also highly non-trivial. First, we must show that the stochastic system possesses a unique stationary distribution. Then, after showing tightness of the considered sequence, we must characterize the limit of all converging subsequences, and prove that they all coincide, having the form in (2). We note that it is not clear that there necessarily exists a unique stationary point for the fluid limit, as there may be more than one stationary point for the fluid limit when the service and patience times are independent; see the discussion above Lemma 3.1 in [6].

Acknowledgements Ohad Perry was partially supported by NSF Grant CMMI-2006350.

References

1. BizReport. Americans abandon purchases in-store after 8 minutes waiting in line. http://www.bizreport.com/2014/03/americans-abandon-purchases-in-store-after-8-minutes-waiting.html, (2014)
2. Chan, C.W., Farias, V.F., Escobar, G.J.: The impact of delays on service times in the intensive care unit. Manag. Sci. **63**(7), 2049–2072 (2017)
3. De Vries, J., Roy, D., De Koster, R.: Worth the wait? How restaurant waiting time influences customer behavior and revenue. J. Op. Manag. **63**, 59–78 (2018)
4. Decreusefond, L., Moyal, P.: A functional central limit theorem for the M/GI/∞ queue. Ann. Appl. Probab. **18**(6), 2156–2178 (2008)
5. Kang, W., Ramanan, K.: Fluid limits of many-server queues with reneging. Ann. Appl. Probab. **20**(6), 2204–2260 (2010)
6. Kang, W., Ramanan, K.: Asymptotic approximations for stationary distributions of many-server queues with abandonment. Ann. Appl. Probab. **22**(2), 477–521 (2012)
7. Kaspi, H., Ramanan, K.: Law of large numbers limits for many-server queues. Ann. Appl. Probab. **21**(1), 33–114 (2011)
8. Reich, M.: *The Offered-Load Process: Modeling, Inference and Applications*. PhD thesis, Technion-Israel Institute of Technology, Faculty of Industrial Engineering and Management, (2012)
9. Wu, C., Bassamboo, A., Perry, O.: Service system with dependent service and patience times. Manag. Sci. **65**(3), 1151–1172 (2019)
10. Wu, C.A.: Queueing Models for Service Systems with Dependencies. PhD thesis, Northwestern University, (2018)
11. Wu, C.A., Bassamboo, A., Perry, O.: A unified fluid model for service systems with exogenous and endogenous dependencies. *Working paper*, (2022)
12. Wu, C.A., Bassamboo, A., Perry, O.: When service times depend on customers' delays: A solution to two empirical challenges. *Operations Research, forthcoming*, (2022)
13. Yu, L., Perry, O.: Many-server heavy-traffic limits for queueing systems with perfectly correlated service and patience times. arXiv preprint arXiv:2008.12890, (2022)