Poster: Skewness Variance Approximation for Dynamic Rate Multi-Server Queues with Abandonment

William A. Massey and Jamol Pender
Department of Operations Research and Financial Engineering
Princeton University
{wmassey, jpender}@princeton.edu

ABSTRACT

Large scale systems such as customer contact centers, like telephone call centers, as well as healthcare centers, like hospitals, have customer inflow-outflow dynamics with many common features. The customer arrival patterns may have time of day or seasonal effects. Moreover, customer population sizes tend to be large where the individual actions are intrinsic and independent of other customer actions and there are multiple service agents, so many customers have access to services in parallel. Finally, arriving customers engaging in service may be delayed if all the available agents are busy. Moreover, these waiting customers may decide to leave the systems if they feel that their delay in receiving service is excessively long.

A fundamental Markov process, with dynamic rates, queueing model for large service systems is a multi-server queue with non-homogeneous Poisson arrivals as well as service and customer abandonment times that are exponentially distributed. An asymptotic scaling for these queues leads to both functional strong law of large numbers and central limit theorems. The first yields a dynamical system that we call a *fluid model*. The second scaled limit yields a *diffusion process model* that is Gaussian under conditions involving the fluid model not lingering too closely to the number of servers. Finally, the fluid model coupled with the mean and covariance of the Gaussian model is also a dynamical system. These results are a special case of a general asymptotic theory for *Markovian service networks* as derived in Mandelbaum, Massey, and Reiman [1].

In practice, these results yield useful Gaussian approximations to the transient queue length distribution. However, these estimates tend not to work as well when the lingering effect is significant. We can improve these methods by introducing a new technique that is called the *Gaussian-skewness approximation* (GSA). It is the special case of a general Hermite polynomial expansion for a Gaussian random variable. We then obtain dynamical systems that approximate the mean, variance and higher cumulant moments of the queueing process for a more accurate, non-Gaussian estimation.

The numerical example that we consider in this paper, to illustrate our approximation methods, has a time interval $(0, 20]$, an arrival rate function $\lambda(t) = 20 + 10\sin(t)$, a service rate $\mu = 1.0$, abandonment rate $\beta = 0.5$, and the number of servers is $c = 20$. We highlight the specific time $t = 4.25$ in Figure 1 since the simulated histogram for the queue length distribution here is when the mean number in the queueing system has gone from being greater than the number of servers, or overloaded, to equaling the number exactly at this time. This is when our queue length is the least Gaussian in appearance. Note that this is also a time where the skewness attains a locally maximal value. Figure 1 also illustrates that GSA (solid curve) skews the its density distribution more to the left compared to the other two Gaussian approximations (both dotted lines where the smaller one at the peak, which also is the largest one left of the peak, uses the fluid model for the mean at this time and the diffusion model for the variance). Moreover, GSA yields an asymmetric density so it does a better job of fitting both queueing distributional tails here than any Gaussian method.

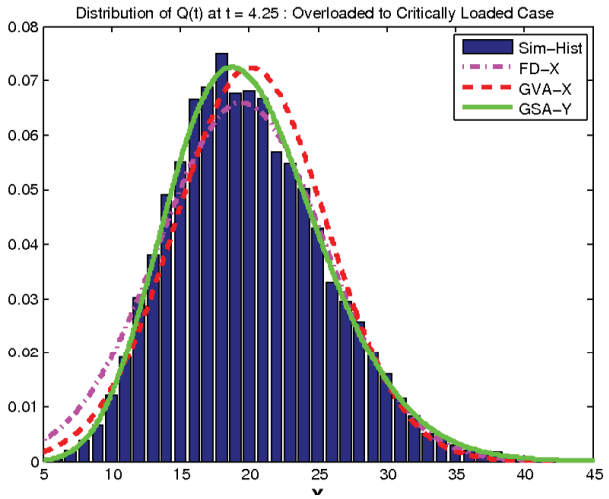

Figure 1: Comparison of Fluid/Diffusion, GVA, and GSA Densities to Histograms of Simulated Queue Length Distributions.

References

[1] Mandelbaum, A., Massey, W. A., Reiman, M. Strong Approximations for Markovian Service Networks. *Queueing Systems*, 30 (1998) pp. 149–201.