Limit Theorems for Markov-modulated Queues

H. Thorsdottir

Summary of *Limit theorems for Markov-modulated queues*

Halldora Thorsdottir

April 15, 2016

This thesis considers queueing systems affected by a random environment, with the primary focus being that of evaluating the performance of these queues in specific asymptotic regimes. Embedding a queueing system in a random environment is a way to add flexibility to a model. This flexibility comes at the cost of increased complexity, in that the queue becomes a doubly random system, i.e. the already stochastic arrival and service processes are also assumed to have randomly fluctuating parameters governed by the external environment. In the scope of this thesis the environment can in theory represent anything that can be modelled as a random process on a discrete state space, such as weather dynamics or the state of the economy. We frequently refer to the environment as the background process or modulating process. In addition to an introduction, the thesis consists of four chapters; Chapters 2 to 5 are based on journal papers that have been published. Chapters 2, 3 and 4 all study infinite server systems, whereas in Chapter 5 a single server queue is studied.

Chapters 2, 3 and 4 all exploit the concept of time-scale separation. The scaling of choice is applied to both the environment and the arrival process. Both are pushed to infinity albeit at different speeds, imposing a central limit theorem (CLT) type of scaling. On the one hand, when the environment is sped up more than the arrival process, it will only be perceived as an average from the perspective of the main process, the queue length. Instead of having multiple arrival and service rates due to the modulation, in the limit one effectively only observes an average arrival and service rate, which greatly simplifies the analysis. On the other hand, slowing down the environment relative to the arrivals, as in Chapter 4, yields a sequence of temporary steady-states. In that case the deviation between the transient and the equilibrium distribution of the environment, expressed in terms of the so-called deviation matrix for Markov chains, plays a crucial role.

Under the CLT scaling, Gaussian limits are derived. In Chapters 2 and 3, which contain results at the transient level, the limiting distribution is identified as the normal distribution, whereas the functional CLT of Chapter 4 results in a limiting Gaussian process of the Ornstein-Uhlenbeck (OU) type. The heavy-traffic scaling in Chapter 5 lets the arrival rate be increased such that the traffic intensity reaches its critical point. Typically the scaling yields reflected Brownian motion limits; here its stationary counterpart, the exponential distribution, is obtained as the limiting distribution of the steady-state workload and queue length.

The methodology of Chapters 2, 3 and 5 is likely familiar to the reader of

queueing literature. We set off with fixed-point equations to describe the infinite server queue in Chapters 2 and 3, and balance equations for the M/G/1 queue in Chapter 5. Due to the modulating background process the equations become particularly uninviting. By applying the right scaling and Taylor expansion, the equations simplify considerably, which helps in deriving the limits. The work in Chapter 4 is derived under a different framework and methodology, primarily based on unit-rate Poisson processes and the martingale CLT. The main advantage of this toolkit is that it yields a functional limiting result, whereas the methods of the other chapters yield finite-dimensional convergence.

While the main topic of this thesis is the behaviour of modulated queues in particular scaling regimes, it contains specific results for non-scaled processes as well, presented in Chapters 3 and 5. These results, which are obtained using transforms, are primarily in terms of recursions for moments and differential equations that describe properties of the distribution of the number of customers and workload.

We summarize the main results of the thesis. After speeding up the environment that modulates the Poisson arrivals to an infinite server queue, the arrival process is shown to be asymptotically Poisson with a uniform rate, see Chapter 2. By also speeding up the arrival rates, the scaled and centered queue length converges to a normally distributed random variable. Here the background process has deterministic transition times, yielding a semi-Markov-modulated system. These results are extended in Chapter 3 to a multi-dimensional CLT for an M/G/∞ queue under Markov-modulation. Chapter 4 contains a functional CLT for the queue length process with Markov-modulated arrivals and nonmodulated, exponential service times. The martingale CLT is applied to prove weak convergence to an OU process, where the environment moves either faster or slower than the arrival process. In Chapter 5, assuming generally distributed service requirements and a fairly general class of service disciplines, it is shown that the workload of an M/G/1 queue with Markov-modulated service capacity converges to an exponentially distributed random variable in heavy traffic. The discriminatory processor sharing service discipline is applied to the case of exponentially distributed service requirements and multiple customer classes. The queue length vector for the various classes undergoes a state-space collapse in the limit of heavy-traffic scaling, also giving the exponential distribution.