Analysis of structured multi-dimensional Markov processes

Jori Selen

Markov processes provide an essential instrument for modeling and analyzing a large variety of systems and networks, including manufacturing systems, communication networks, traffic networks and service systems such as clinics or hospitals. These processes are detailed enough to capture the essential system dynamics, but are simple enough in terms of mathematical structure to be amenable for theoretical analysis.

Markov processes fall under the umbrella of stochastics, the branch of mathematics that aims to establish rigorous statements about systems that are inherently uncertain, and therefore subject to some degree of randomness. A classical example is a queue, in which jobs need to wait for service. The queue grows when new jobs arrive and shrinks when jobs complete service. Queues occur virtually everywhere and can be seen as stochastic systems subject to variability. Under certain assumptions, a queueing system can be modeled as a Markov process.

In this thesis we are primarily interested in determining the equilibrium distribution of Markov processes that describe the queueing dynamics. The equilibrium distribution characterizes the long-term fraction of time that a Markov process spends in each of the possible states. Without using any further structural properties, finding the equilibrium distribution requires solving the balance equations. This linear system of equations is generally difficult to solve, especially when the state space is countably infinite. Nonetheless, in this thesis we develop and extend methods for determining the equilibrium distribution by exploiting the structure that is governed by the interaction between states.

We consider three classes of Markov processes that each share structural properties. The first class of processes take values in a two-dimensional state space, where one dimension is infinite and the other finite. This class is motivated by multi-server queueing systems with customers that require special service when delayed; heterogeneous servers; or batch services or arrivals. To determine the equilibrium distribution of this class of models, we (1) use the matrix-geometric method in combination with a novel technique for determining the boundary probabilities; and (2) extend the spectral expansion method to analyze the heterogeneous servers case.

The second class is inspired by priority queueing systems. The dimensions of the associated Markov processes represent the number of customers of each class in the system, which leads to a multi-dimensional state description. We present a novel recursive method, based on the matrix-analytic method, to determine the equilibrium distribution of a single-server queueing system with an arbitrary number of priority classes and class-dependent service rates. For a multi-server priority system with two classes, we develop a method that determines the Laplace transforms of the transition functions of the associated Markov process. This method extends the clearing analysis on phases method and the matrix-analytic method to the complex domain and combines these techniques to develop a recursion.

Finally, the third class extends the classical join-the-shortest-queue model and the techniques associated with this model. In particular, we analyze a queueing system with two single-server queues with different service rates. Arriving jobs are routed according to the shortest-expected-delay routing protocol, which aims to balance the sum of the expected service times of all jobs at each server. The

associated Markov process is three-dimensional and we extend the compensation approach---which is used for the join-the-shortest-queue model---to determine the equilibrium distribution. This approach leads to series expressions for the equilibrium probabilities in terms a countably infinite number of product-form solutions. For a related model with generally distributed service times and a processor-sharing service discipline, we use the above results in combination with a single-queue approximation to approximate key performance measures.