Gennady Shaikhet – Research Statement

Work: 412-268-3781 • shaikhet@cmu.edu • www.math.cmu.edu/~shaikhet

My main research interests lie in the applications of probability theory, with a particular focus on operations research, queueing theory, approximations and control of stochastic processes and mathematical finance. My other research interests include optimization, simulation, combinatorics and graph theory. Below, I first describe my past work and then briefly outline some directions of future research.

Past work

Queueing Theory

Background: Multi-server queues arise as models in many applications, including telephone call centers, computer data systems and healthcare systems. Multi-server queueing systems are typically harder to analyze than single-server systems. An exact analysis is in many cases intractable, but valuable insight can often be obtained via suitable asymptotic approximations. In the applications described above, there are typically a large number of servers and the probability of a positive wait time for an arriving customer lies strictly between 0 and 1. It is therefore natural to consider the asymptotic limit of these systems, as both the arrival rate and the number of servers go to infinity in such a way that the (limiting) probability of wait remains non-trivial. This regime is different from the classical so-called "heavy traffic" regime, and poses several new challenges. The regime was formalized in [6], in the context of a single multi-server queue by Halfin and Whitt. They considered the performance of the asymptotic limit of such system under the assumption of exponential service times. However, in many applications statistical analysis has shown that service times are typically not exponentially distributed. Moreover, in addition to performance analysis, it is also of interest to design optimal or near-optimal controls for these systems.

Control: Consider a queueing system with I customer classes and J server pools. Customer arrivals for each class follow a renewal process and each pool has many statistically identical, exponential servers. These servers work independently, offering service to different classes of customers at rates, which depend on both the class i and the pool j. Customers may abandon while waiting to be served, according to exponential clocks with class dependent. We consider a routing and scheduling control problem in which customers are to be routed to pools in a way that performance measures, such as average queue-lengths, are minimized. In the Halfin – Whitt

regime, those problems are formally equivalent to multidimensional diffusion control problems, for which it is typically hard to obtain explicit solutions. In [3] we show, that in the case, when the service rates depend only on the pool, the optimal solution can remarkable be obtained by solving a certain one-dimensional diffusion control problem. Such "dimensionality reduction" techniques make the analysis of high-dimensional systems numerically tractable and are extremely important for applications.

Null controllability: In the paper [5] we have discovered and studied a surprising and very unusual heavy traffic phenomenon, when, under certain control policy, a critically loaded system operates with empty queues, as if it is underloaded. The phenomenon was called null controllability. A sequel [4] to that work, that mostly uses graph theory and combinatorial optimization, investigates conditions on the graph of the queueing network, as well as the routing policy, that allow null controllability to happen.

Optimization: In more recent work [2], we relax the exponential assumption on the service distribution. We aim to find a control policy that minimizes the long run average holding cost. We show that under the fluid scaling and overload condition the problem is analogous to solving a certain deterministic problem, the solution of which suggests an optimal routing strategy.

Finance

Transaction costs: I was first exposed to research in finance during my Master's thesis. Optimal trading strategies in finance have been well-studied in the absence of transaction costs. However, to understand real-world systems, it is important to incorporate transaction costs. Together with Gady Zohar, we studied the problem of optimal trading in the presence of fixed transaction costs. In theory, such problems may be treated by methods of impulse control, unfortunately with little hope for an explicit solution. In that work we considered a method to find an approximately optimal solution for the case of small transaction costs.

Market Microstructure: Recently, electronic platforms on many stock exchanges aggregate all outstanding limit orders in a so called limit order book, that is available to market participants; and market orders are executed against the best available prices. In [1], together with Prof. Steven Shreve and PhD student Silviu Predoiu we have constructed an optimal execution strategy for a large purchase of an underlying asset over a fixed interval of time.

Future plans

Queueing Theory

Non-exponential service times: Until recently, the exponential assumption on the service time distribution was key in many works. The reason is that the memoryless property of the exponential distribution results in some remarkable simplifications in the representation of these processes. There are very few works dealing with performance measures of single class multi-server queueing systems in some particular non-exponential setting and there are almost no works that deal with control of queueing systems with several customer classes and several service stations, in heavy traffic, in the non-exponential case. I plan to build on the measure-valued process approach introduced by Kaspi and Ramanan [7] to consider such control problems. My results should provide good heuristics for near-optimal routing in many-server systems, which can be further optimized / fine-tuned with the help of simulations. In addition, they would also lead to some theoretical developments in the study. I would also intend to investigate many-server queueing systems that arise in health-care systems. The modeling assumptions here are likely to be somewhat different from the call center setting, and I plan to study how the methods developed above would have to be adapted to analyze such systems.

Phase-Type service times: There are still many interesting challenges, both in approximations and control of queueing systems, where the service distribution is exponential, or a close relative to it, like phase-type. I plan to generalize some of my previous results, especially null controllability, to the case of phase-type service distribution. This may lead to a deeper insight into the phenomenon, as well to some theoretical contribution to the field.

Numerical methods: Even for the exponential service times there are only few cases when the exact structure of the optimal control policy is known. Such explicit solutions will contribute to more practical aspects of the subject. I plan to use numerical methods and simulations to gain more insight on the optimal control.

Finance

Market Microstructure: This is a new emerging area of considerable importance that lies in the interface of queueing theory and finance. I plan to continue working in that direction.

Discussed papers

1. **Optimal Execution in a General One-Sided Limit-Order Book** (2009). (with Steven Shreve and Silviu Predoiu). Submitted.
2. **A Fluid Control Problem in Overloaded Queueing Networks with General Service Times** (2010). Submitted.
3. **Simplified control problems for multi-class many-server queueing systems** (2009). (with Rami Atar and Avishai Mandelbaum). *Math. Oper. Res. Vol. 34, No. 4, November 2009, pp. 795-812.*
4. **Critically loaded queueing models that are throughput sub-optimal** (with Rami Atar). *Ann. Appl. Prob., 2009, Vol. 19, No. 2, 521–555.*
5. **Queueing Systems with Many Servers: Null Controllability in Heavy Traffic** (with Rami Atar and Avishai Mandelbaum). *Ann. Appl. Prob., 2007, Vol. 16, No. 4, 1764–1804.*
6. Halfin S. and Whitt W. (1981). **Heavy traffic limits for queues with many exponential servers.** *Oper. Res. 29, No. 3. 567-588.*
7. Kaspi H. and Ramanan K. (2007). **Law of large numbers limits for many-server queues.** Preprint