Syllabus for IEOR 8100: PhD Seminar on Queueing Theory

Ward Whitt

Department of Industrial Engineering and Operations Research, Columbia University,
New York, NY, 10027; ww2040@columbia.edu

September 4, 2016

1 Course Overview

This will be a research course in which the participants will read research papers, conduct indivdual research projects and make presentations to the class. Students already engaged in research can present their own recent research results as well as important background papers and new research papers in the same area. Other students can present new research as well as research papers. There will be no tests or homework problem sets.

Consistent with my recent research and the research of my current students, the main themes for the seminar will be: (i) heavy-traffic approximations based on stochastic-process limits, (ii) robust queueing, (iii) queues with time-varying arrival rates, (iv) fitting queueing models to system data and (v) applications of queueing models to service systems. The following sections point out relative background and recent research efforts in these areas. The papers provide links to the previous literature. (My book [45] and almost all of my papers are available for downloading from my web page. For the full book, use the link: http://www.columbia.edu/∼ww2040/jumps.html)

Students may also pursue a research topic of their own choosing. Former student Song-Hee Kim started her research interest in healthcare by electing to study queueing models in healthcare in a similar seminar a few years ago. Queueing theory can be usefully applied to many problems. To illustrate one nonstandard application, I refer to my papers on managing the pace of play in golf [7, 8, 15, 50].

The seminar will be run from the instructor's web page: http://www.columbia.edu/∼ww2040

2 Probability Foundations

2.1 Background on Probability, Stochastic Processes and Queueing Theory

The seminar assumes a background roughly equivalent to completing the IEOR first-year core doctoral course on stochastic models, IEOR 6711 and 6712. A good (relatively advanced) textbook at this level is Asmussen [2]. Understanding of the basics about stochastic processes and queueing models will be assumed, but could be picked up along the way, given the appropriate mathematical experience. Among the good textbook introductions to queueing theory are Cooper [9], Kleinrock [22, 27] and Wolff [58].

2.2 Heavy-Traffic and Stochastic-Process Limits

Queueing models can often be better understood and approximated with the aid of heavy-traffic limits. These heavy-traffic limits draw on the theory of convergence of measures on function spaces or, equivalently, the theory of convergence of stochastic processes, as discussed in the textbooks by Billingsley [5, 6]. An overview of that theory without many proofs appears in my own [45]. The book [45] focuses on applications of that theory to establish heavy-traffic limits for queueing models. The book focuses on conventional heavy-traffic limits in which a finite number of servers is held fixed while the traffic intensity is allowed to increase toward its critical value. Another useful book that focuses more on the reflected Brownian motion (RBM) limit process is Harrison [19]. A book on similar topics from a more applied point of view is Newell [38].

Recently I have been focusing on many-server heavy-traffic limits, in which both the arrival rate and the the number of servers increase without bound. An early paper on that is [18]. A survey paper on the martingale methods often used to establish many-server heavy-traffic limits is [39].

3 Robust Queueing

Robust queueing is a new approach to develop useful approximations for hard queueieng problems, initiated by Bertsimas and his students [4]. The key idea is to develop bounds and approximations by replacing a hard stochastic problem by an optimization problem that is easier to analyze. Currrent student Wei You and I have begun trying to contribute to this area [54, 55].

4 Time-Varying Queues

Just like the heavy-traffic limits, the literature on queues with time-varying arrival rates divides into two main categories: (i) queues with many servers and (ii) queues few servers, e.g., only one.

4.1 Many-Server Queues

Many-server queues tend to be easier to analyze than single-server queues, especially if we approximate them by infinite-server queues. There is now a quite large literature on infinite-server queues with time-varying arrival rates, largely stemming from my earlier papers with W. A. Massey [12, 13, 34] and continuing with more recent two-parameter heavy-traffic limits with former students Guodong Pang [40] and Yunan Liu [1, 30].

4.1.1 Staffing to Stabilize Performance

For many-server queues, an important problem is staffing (choosing the time-varying number of servers) to stabilize performance. This line of research began with my paper with Jennings, Mandelbaum and Massey [21] and continued with [14] and then [20, 28, 29] with students Andrew Li (now at MIT), Yunan Liu (now at NCSU) and Jingtong Zhao,. These use the modified-offered-load (MOL) approach, which is reviewed in the short survey papers [46, 48].

4.1.2 The Time-Varying Little's Law

Little's law (LL, $L = \lambda W$) is intimately connected to infinite-server queues; see p. 238 of my review paper [44]. Background on LL and the time-varying gneralization (TVLL) can be found in my survey paper [44] and in [25, 26] with former student Song-Hee Kim. Research on the time-varying Little's law is now being conducted with current student Xiaopei Zhang [57]. It emerged from analyzing data from an Isreali emergency department [56].

4.2 Single-Server Queues

4.2.1 Understanding the Performance

For developing a fundamental understanding of the behavior of time-varying queues, the seminal papers were by Newell [35, 36, 37]. Heavy-traffic limits have been established by Mandelbaum and Massey [33] and in my more recent short contributions [49, 53].

4.2.2 Time-Varying Robust Queueing

Time-varying robust queueing (TVRQ) is being studied with current student Wei You [54]. A robust queueing approach to the transient behavior of stationary queues is proposed by [3].

4.2.3 Staffing to Stabilize Performance

A service-rate control used to stabilize performance in a single-server queue is studied in my recent [51].

4.2.4 Simulation Methods

Current student Ni Ma has been studying and applying simulation methods, including a rare-event method, to study single-server queues with time-varying arrivals [31, 32].

5 Fitting Queueing Models to Service System Data

A major focus of my recent work has been on methods to fit queueing models to service system data. Examples are [23, 24] with Song-Hee Kim and [56] with current student Xiaopei Zhang. A new direction is fitting birth-and-death processes to the sample path of a queue length process. This is discussed in [10, 11, 47, 52], the second and third with former IEOR undergraduate James Dong (now in the graduate progem at Cornell).

6 Applications to Service Systems

Much of my recent research has been motivated by applications to service systems, such as call centers and healthcare systems.

6.1 Assigning Arrivals to Servers

Former students Itai Gurvich (now at Cornell NY) and Ohad Perry (now at Northwestern University) worked on rules for assigning arriving customers to servers, e.g., [16, 17] and [41, 42]. Current student Xu Sun has also been working in this area. We have developed a way to use routing rules to create work breaks for agents from available idleness [43].

References

[1] Aras, K., Liu, Y. and Whitt, W. (2014). Heavy-traffic limit for the initial content process. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[2] Asmussen, S. (2003). *Applied Probability and Queues*. New York: Springer, 2nd edition.

[3] Bandi, C., Bertsimas, D. and Youssef, N. (2014). Robust transient multi-server queues and feedforward networks. Unpublished manuscript, MIT ORC Center.

[4] Bandi, C., Bertsimas, D. and Youssef, N. (2015). Robust queueing theory. *Operations Research* 63(3):676–700.

[5] Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley, 1st edition.

[6] Billingsley, P. (1999). *Convergence of Probability Measures*. New York: Wiley.

[7] Choi, M. and Whitt, W. (2016). The advantage of the wave-up rule in golf: a simulation study. In preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[8] Choi, M. and Whitt, W. (2016). The impact of slow groups in golf: a simulation study. In preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[9] Cohen, J. W. (1982). *The Single Server Queue*. Amsterdam: North-Holland, 2nd edition.

[10] Dong, J. and Whitt, W. (2015). Stochastic grey-box modeling of queueing systems: Fitting birth-and-death processes to data. *Queueing Systems* 79:391–426.

[11] Duffield, N. G. and Whitt, W. (2015). Using a birth-and-death process to estimate the steady-state distribution of a periodic queue. *Naval Research Logistics* 62:664–685.

[12] Eick, S. G., Massey, W. A. and Whitt, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci* 39:241–252.

[13] Eick, S. G., Massey, W. A. and Whitt, W. (1993). The physics of the $M_t/G/\infty$ queue. *Oper Res* 41:731–742.

[14] Feldman, Z., Mandelbaum, A., Massey, W. A. and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Sci* 54(2):324–338.

[15] Fu, Q. and Whtt, W. (2015). Analyzing the pace of play in golf. *Journal of Sports Analytics* 1(1):43–64.

[16] Gurvich, I. and Whitt, W. (2009). Queue-and-idleness-ratio controls in many-server service systems. *Mathematics of Operations Research* 34:363–396.

[17] Gurvich, I. and Whitt, W. (2009). Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Oper Management* 11:237–253.

[18] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.

[19] Harrison, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*. New York: Wiley.

[20] He, B., Liu, Y. and Whitt, W. (2016). Staffing a service system with non-poisson nonstationary arrivals. *Probability in the Engineering and Informational Sciences* .

[21] Jennings, O. B., Mandelbaum, A., Massey, W. A. and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Sci* 42:1383–1394.

[22] Keiding, N. (1975). Maximum likelihood estimation in the birth-and-death process. *Ann Statist* 3:363–372.

[23] Kim, S. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Oper Management* 16(3):464–480.

[24] Kim, S. and Whitt, W. (2014). Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics* 17:307–318.

[25] Kim, S.-H. and Whitt, W. (2013). Estimating waiting times with the time-varying Little's law. *Probability in the Engineering and Informational Sciences* 27:471–506.

[26] Kim, S.-H. and Whitt, W. (2013). Statistical analysis with Little's law. *Operations Research* 61(4):1030–1045.

[27] Kirstein, B. M. (1976). Monotonicity and comparability of time-homogeneous markov processes with discrete state space. *Math Operationsforsch Statist* 7:151–168.

[28] Li, A., Whitt, W. and Zhao, J. (2016). Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* 30:185–211.

[29] Liu, Y. and Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper Res* 60(6):1551–1564.

[30] Liu, Y. and Whitt, W. (2014). Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability* 24(1):378–421.

[31] Ma, N. and Whitt, W. (2016). Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters* 102:202–207.

[32] Ma, N. and Whitt, W. (2016). A performance algorithm for periodic queues. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[33] Mandelbaum, A. and Massey, W. A. (1995). Strong approximations for time-dependent queues. *Mathematics of Operations Research* 20(1):33–64.

[34] Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13(1):183–250.

[35] Newell, G. F. (1968). Queues with time dependent arrival rates, I: the transition through saturation. *Journal of Applied Probability* 5:436–451.

[36] Newell, G. F. (1968). Queues with time dependent arrival rates, II: the maximum queue and the return to equilibrium. *Journal of Applied Probability* 5:579–590.

[37] Newell, G. F. (1968). Queues with time dependent arrival rates, III: a mild rush hour. *Journal of Applied Probability* 5:591–606.

[38] Newell, G. F. (1982). *Applications of Queueing Theory*. London: Chapman and Hall, 2nd edition.

[39] Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* 4:193–267.

[40] Pang, G. and Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* 65:325–364.

[41] Perry, O. and Whitt, W. (2009). Responding to unexpected overloads in large-scale service systems. *Management Science* 55:1353–1367.

[42] Perry, O. and Whitt, W. (2013). A fluid limit for an overloaded x model via a stochastic averaging principle. *Mathematics of Operations Research* 38:294–349.

[43] Sun, X. and Whitt, W. (2016). Creating work breaks from available idleness. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[44] Whitt, W. (1991). A review of $L = \lambda W$. *Queueing Systems* 9:235–268.

[45] Whitt, W. (2002). *Stochastic-Process Limits*. New York: Springer.

[46] Whitt, W. (2007). What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics* 54(5):476–484.

[47] Whitt, W. (2012). Fitting birth-and-death queueing models to data. *Statistics and Probability Letters* 82:998–1004.

[48] Whitt, W. (2013). Offered load analysis for staffing. *Manufacturing and Service Operations Management,* 15(2):166–169.

[49] Whitt, W. (2014). Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters* 42:458–461.

[50] Whitt, W. (2015). The maximum throughput on a golf course. *Production and Operations Management* 24(5):685–703.

[51] Whitt, W. (2015). Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* 81:341–378.

[52] Whitt, W. (2016). Heavy-traffic fluid limits for periodic infinite-server queues. *Queueing Systems* 80.

[53] Whitt, W. (2016). Heavy-traffic limits for a single-server queue leading up to a critical point. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[54] Whitt, W. and You, W. (2016). Time-varying robust queueing. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[55] Whitt, W. and You, W. (2016). Using robust queueing to expose the impact of dependence in single-server queues. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[56] Whitt, W. and Zhang, X. (2015). A data-generated queueing model of an emergency department. In preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[57] Whitt, W. and Zhang, X. (2016). Periodic little's law in discrete time. In preparation, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[58] Wolfe, R. W. (1989). *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice-Hall.