Fluid Limit Approximations of Stochastic Networks

Remerova, M.

2014

The drawback of this approach is that it is restricted to the cases when there exist explicit formulas for the generators. On the other hand, if the exact formulas are known, the proof of the RHS of (1.10) is usually straightforward (often via Taylor expansion). For more details on the theory behind the method, we refer to Ethier and Kurtz [40]; for examples of its application, see [37, 76, 89].

1.6 Overview of the thesis

Here we provide a summary of the subsequent chapters of this thesis. They all develop fluid limit approximations, but the models they consider are very different and thus the analysis techniques vary greatly from one chapter to another. Each chapter corresponds to a paper.

- Chapter 2 is based on [43] M. Frolkova, S. Foss, and B. Zwart. Fluid limits for an ALOHA-type model with impatient customers. *Queueing Systems*, 72:69–101, 2012.
- Chapter 3 is based on [92] M. Remerova, J. Reed, and B. Zwart. Fluid limits for bandwidth-sharing networks with rate constraints. Accepted for publication in *Mathematics of Operations Research*, 2013.
- Chapter 4 is based on [93] M. Remerova, S. Foss, and B. Zwart. Random fluid limit of an overloaded polling model. *Advances in Applied Probability*, 46:76–101, 2014.
- Chapter 5 is based on [91] M. Remerova and B. Zwart. Fluid limits of a PS-queue with multistage service. In preparation, 2013.

Chapter 2: An ALOHA-type model with impatient customers ALOHA protocols are random multiple-access protocols. They are designed for networks with star configurations where multiple client nodes talk to the hub node at the same frequency. Consequently, if there are two or more client nodes talking simultaneously, they are all in conflict, preventing each other from being heard by the hub. The common idea of ALOHA protocols is “try to send your data and, if your message collides with another transmission, try resending later”.

We study a generalisation of the conventional centralised time-slotted ALOHA model where impatience of users is allowed, which we assume to be caused by the overload regime. We apply to the (multidimensional) population process a time-space fluid scaling that lets users become more and more patient. Our first result is a description of fluid limits as solutions to a system of deterministic differential equations. The most challenging part of the proof of this result is to eliminate problems at zero. They arise because of the centralised protocol, which assumes that each time slot each user tries to transmit with probability one over the total population. The second main result of Chapter 2 is convergence of fluid limits over time to the unique fixed point. We prove it by means of a Lyapunov function.

Chapter 3: Bandwidth-sharing networks with rate constraints In a bandwidth-sharing network, elastic flows compete for service on several links. Link capacities are redistributed among the flows as their population changes, and bandwidth allocations are chosen in such a way that the network utility is always maximised. This setting was introduced by Massoulié & Roberts [97, 75], and nowadays is considered classical.

In Chapter 3, we modify the classical bandwidth-sharing setting by imposing constraints on processing rates of individual flows. We study the behavior of the model under the large capacity scaling, and with that we mean that the rate constraints, flow sizes and their patience times remain of a fixed order while the network capacity and arrival rates grow large. Note that this scenario is standard in practice. Under general stochastic assumptions, we characterise fluid limits of a process that contains full information about the system state, including residual flow sizes and their residual patience times. In particular, we extend the fluid limit result of Reed and Zwart [88] for Markovian stochastic assumptions. We also prove a new type of result for bandwidth-sharing networks: convergence of the network stationary distribution to the fixed point of the fluid limit equations under the fluid scaling (we need stricter stochastic assumptions here). Moreover, we show that, in many cases, the fixed point solves a strictly concave optimization problem, and thus can be computed in a polynomial time, which is a surprisingly efficient way to approximate such a complicated stochastic model.

Chapter 4: Random fluid limit of an overloaded polling model For many basic queueing systems, fluid limits are deterministic functions. In Chapter 4, we study a cyclic polling model under conditions that lead to a random fluid limit. These conditions are zero initial state and overload. We allow a wide class of service disciplines, which we call "multigated" and which assume a random number of iterations of conventional gated service. Exhaustive policy is in this class as well. Such disciplines ensure that the system population evolves as a multitype branching process. This provides us with our main tool — the Kesten-Stigum theorem that characterises long-time behavior of supercritical (or "overloaded") branching processes. The scaling regime we apply in this chapter is simply zooming out, and the fluid limit we obtain has a rather interesting structure. Firstly, the fluid limit oscillates frequently often in the neighborhood of zero. Secondly, all its trajectories can be mapped by a linear time-space scaling into the same deterministic function. An additional contribution of this chapter is that we develop a method of proving finiteness of moments of the busy period in an $M/G/1$ queue. It is inspired by a technical moment condition of the Kesten-Stigum theorem.

Chapter 5: A processor-sharing queue with multistage service This chapter considers a Markovian processor-sharing queue where service of each customer consists of several stages with independent service requirements. As in Chapter 2, the system is overloaded and customers are impatient. We develop fluid limit approximations of the per-stage population process and characterise them as solutions to a system of deterministic differential equations. We also establish relations between our fluid limits and measure-valued fluid limits in Gromoll et al. [47] for a processor-sharing queue with single-stage service. This allows us to prove that all fluid limits stabilise to the unique fixed point over time. Additionally, we discuss Lyapunov functions for processor sharing. Existence of Lyapunov functions for models that combine impatience and routing (note that multistage service corresponds to tandem routing) is an open problem. We suggest partial solutions.

1.7 Notation

Here we list the notations common for all of the subsequent chapters.

To define x as equal to y, we write $x := y$ or $y =: x$. We abbreviate the left-hand side and right-hand side of an equation as "LHS" and "RHS", respectively.

The standard sets are: the natural numbers $\mathbb{N} := \{1, 2, \ldots\}$, integers $\mathbb{Z} := \{0, \pm 1, \pm 2, \ldots\}$ and non-negative integers $\mathbb{Z}_+ = \{0\} \cup \mathbb{N}$, the real line $\mathbb{R} := (-\infty, \infty)$ and non-negative half-line $\mathbb{R}_+ := [0, \infty)$.

By e we denote the base of the natural logarithm.

The following operations are defined on $x, y \in \mathbb{R}$:

$$\begin{aligned} &x \vee y := \max\{x, y\}, \quad x \wedge y := \min\{x, y\}, \\ &x^+ := x \vee 0, \quad x^- := (-x) \vee 0, \\ &\lfloor x \rfloor := \max\{n \in \mathbb{Z} \colon n \le x\}. \end{aligned}$$

The upper limit and lower limit are $\overline{\lim}$ and $\underline{\lim}$, respectively.

All vector notations are boldface. Unless stated otherwise, the coordinates of an I-dimensional vector are denoted by the same symbol (regular instead of bold) with subscripts $1, \ldots, I$ added. Overlining, tildes, sub- and superscripts of vectors remain in their coordinates as well. For example: $\overline{\mathbf{Q}}^r(t) = (\overline{Q}_1^r, \ldots, \overline{Q}_I^r)(t)$, $\boldsymbol{\zeta}^* = (\zeta_1^*, \ldots, \zeta_I^*)$, $\mathbf{L}_i = L_{i,1}, \ldots, L_{i,I}$.

By $\mathbf{0}$ we denote the vector whose coordinates are all zeros, and by $\mathbf{1}$ the vector with all coordinates equal 1. In $\mathbb{R}^I$, we work with two norms: the supremum norm $\|\mathbf{x}\| := \max_{1 \le i \le I} |x_i|$, and the L_1-norm $\|\mathbf{x}\|_1 := \sum_{i=1}^I |x_i|$. The vector inequalities hold coordinate-wise. The coordinate-wise product of vectors of the same dimensionality I is denoted by

$$\mathbf{x} \times \mathbf{y} := (x_1 y_1, \ldots, x_I, y_I).$$

For metric spaces S_1 and S_2, $\mathbf{C}(S_1, S_2)$ stands for the space of continuous functions $f \colon S_1 \to S_2$. For a metric space S, $\mathbf{D}(\mathbb{R}_+, S)$ stands for the space of functions $f \colon \mathbb{R}_+ \to S$ that are right-continuous with left limits. We endow $\mathbf{D}(\mathbb{R}_+, S)$ with the Skorokhod J_1-topology. For a function $f(\cdot)$ defined on (a subset of) $\mathbb{R}$, $f'(t)$ denotes its derivative at t.

The complement of an event E is denoted by $\overline{E}$, and its indicator function by $\mathbb{I}\{E\}$. The indicator function $\mathbb{I}_A(\cdot)$ of an arbitrary set A is defined by $\mathbb{I}_A(x) := \mathbb{I}\{x \in A\}$.