Tel Aviv University

The Raymond and Beverly Sackler Faculty of Exact Sciences
School of Mathematical Sciences
Department of Statistics and Operations Research

Tandem Stochastic Systems: The Asymmetric Simple Inclusion Process

Thesis Submitted to the Senate of Tel-Aviv University in Fulfillment of the Requirements for the Degree of "Doctor of Philosophy" by

Shlomi Reuveni

Prepared under the supervision

of Prof. Uri Yechiali

and Prof. Iddo Eliazar

February 2014

1 Acknowledgments

This thesis concludes a wonderful period I have spent as a Ph.D. student in the department of Statistics and Operations Research at Tel-Aviv university. During this time, I have crossed paths with a few special people to whom I wish to express deep gratitude. First and foremost, I would like to thank my wonderful Ph.D. mentors, professor Uri Yechiali and professor Iddo Eliazar. A warm and caring advisor, who is also wise and knowledgeable, is all one could wish for. I therefore consider myself unimaginably lucky to have been blessed with not one but two such advisors. 'Like father like son' the saying goes and it is thus no wonder that Uri and Iddo, academic father and son, share the same style of heartily mentorship. Under their scepter I have grown and thrived, and words cannot even begin to describe my appreciation and admiration to them. May it be that I myself will get the chance to follow my mentors' footsteps.

I would also like to thank Dr. Tamir Tuller with whom I have collaborated while working on gene translation and the Ribosome Flow Model, and to professor Isaac Meilijson who was my Ph.D. mentor at the time. Although ending up dedicating the bulk of my Ph.D. to a different research topic, my joint work with Tamir and Isaaco has ignited my imagination and piqued my curiosity regarding tandem stochastic systems; without it, this thesis would have never been born. I would like to thank Dr. Ori Hirschberg with whom I have collaborated while working on occupation probabilities and fluctuations in the Asymmetric Simple Inclusion Process. Together, we have had many zealful discussions that made science in its making the joyful, mind provoking experience it's meant to be. I would like to thank my former Ph.D. mentor, professor Yossi Klafter, for allowing me to solidify my mathematical education and pursue a second Ph.D. degree while still being in the midst of the first. To that end, university authorities and anonymous committee members also deserve a word of gratitude. By approving my unusual request—not without deliberations—they have demonstrated the kind of mental flexibility so uniquely identified with the academic spirit.

Writing this thesis would not have been possible without the financial support of several different entities. I would like to thank the Israeli council for higher education for supporting me via its 'converging technologies' program. I would also like to thank my mentor, Uri Yechali, and the school of mathematical sciences for supporting me during the final, and most critical, stages of this work.

I would like to finish with a huge hug to my wife Shira who balances my life and keeps me in check. Last but not least, I would like to thank my Mom, Dad and Sister for long years of nurture and encouragement.

2 Abstract

A tandem stochastic system is a system in which basic stochastic "building blocks" are joined together sequentially to form a more complex stochastic process. For example, think of a sequential assembly line in which the final product of one station is the raw material of the next, or on a process of unidirectional transport in which molecules progress along a narrow pore or a molecular track. This thesis opens with Chapter 3 in which we give a brief introduction to Tandem Stochastic Systems (TSS). In this chapter, we acquaint the reader with two paradigmatic models in the field: the Tandem Jackson Network (TJN) and the Asymmetric Simple Exclusion Process (ASEP). Having emerged independently in the queueing theory and non-equilibrium statistical physics literature, we show that these seemingly unrelated models are actually tightly linked, and further explain how this observation has led us to introduce and explore the Asymmetric Simple Inclusion Process (ASIP) — a model which stands at the heart of this thesis. From a queueing theory perspective, the ASIP is a sequential array of Markovian queues with unbounded capacity and unlimited batch service. From a statistical physics perspective, the ASIP is a model for unidirectional transport with coagulation. The ASIP is analyzed in Chapters 4–8 which are organized as follows.

In Chapter 4, we motivate the study of the ASIP. Combining together probabilistic and Monte-Carlo analysis, we showcase the ASIP's rich statistical complexity — which ranges from 'mild' to 'wild' displays of randomness: Gaussian load and draining, Rayleigh outflow with linear aging, Inverse-Gaussian coalescence, intrinsic power-law scalings and power-law fluctuations and condensation. This chapter is based on publication number [1] in the list of references.

In Chapter 5, we study the dynamics and steady state of the ASIP. We derive evolution equations for the mean and Probability Generating Function (PGF) of the sites' occupancy-vector, obtain explicit results for the above mean at steady state, and describe an iterative scheme for the computation of the PGF at steady state. We further obtain explicit results for the load distribution in steady-state — the load being the total number of particles present in all lattice sites. Finally, we address the problem of load-optimization, and solve it under various criteria. This chapter is based on publication number [2] in the list of references.

In Chapter 6, we explore the ASIP's asymptotic statistical behavior. We consider three different limiting regimes: heavy-traffic regime, large-system regime, and balanced-system regime. In each of these regimes we obtain — analytically and in closed form — stochastic limit laws for five key ASIP observables: traversal time, load, busy period, first occupied site, and draining time. The results obtained yield a detailed limit-laws perspective of the ASIP; numerical simulations demonstrate the applicability of these laws as useful approximations. This chapter is based on publication number [3] in the list of references.

In Chapter 7, we introduce *Catalan's trapezoids*, a combinatorial construct instrumental to the analysis of the ASIP. An iterative scheme for the construction of these trapezoids is presented, and a closed-form formula for the calcu-

lation of their entries is derived. Catalan's trapezoids generalize the renowned Catalan's numbers and the combinatorial interpretation of their entries is discussed in light of Bertrand's famous ballot problem. This chapter is based on publication number [4] in the list of references.

In Chapter 8, we present an exact closed-form expression for occupation probabilities in the ASIP. Our results are expressed in terms of the entries of Catalan's trapezoids. We further prove that the ASIP is asymptotically governed by: (i) an inverse square root law of occupation; (ii) a square root law of fluctuation; and (iii) a Rayleigh law for the distribution of inter-exit times. The universality of these results is discussed. This chapter is based on publication number [5] in the list of references — a result of a fruitful collaboration with Dr. Ori Hirschberg who is to be credited for analyzing the ASIP in the continuum limit.

In Chapter 9, we digress from the main theme of this thesis and discuss the computational modeling of gene translation — a biological process which provides a naturally occurring example for a TSS. Gene translation is a central process in all living organism. This process is however still enigmatic, and contradicting conclusions regarding the essential parameters that determine translation rates, appear in different studies. We introduce the Ribosome Flow Model (RFM), a model which takes into account the stochastic nature of the translation process and the excluded volume interactions between ribosomes. The model is aimed at capturing the effect of codon order, and composition, on the translation process and we demonstrate that, in comparison to commonly used approaches, it gives more accurate predictions of translation rates, protein abundance levels and ribosome densities in several different organisms. This chapter is based on publication number [6] in the list of references and is a result of a fruitful collaboration with Dr. Tamir Tuller.

Contents

3 An Introduction to Tandem Stochastic Systems

Many complex and fundamental processes in nature incorporate a high level of intrinsic randomness. Stochastic events stand at the very bedrock of their micro level description, and the cumulative effect of these events is manifested in their dynamics and functionality. Complex processes may sometimes be viewed as an interconnected network, whose basic building blocks are processes of diminished complexity. Interestingly, even when the isolated behavior of each building block is understood in great detail, the behavior of the aggregate is often extremely hard to predict. This chapter serves as an introduction to Tandem Stochastic Systems (TSS), linear stochastic networks formed by the sequential concatenation of stochastic processing units. The prevalence of TSS throughout the sciences renders this particular class of stochastic networks a special case of interest among a host of scientific communities.

Tandem stochastic systems are systems in which a stochastic input flow (of jobs, molecules, particles, etc.) progresses through a serial array of stochastic processing units. The progress from one processing unit to the consecutive processing unit is governed by a set of rules characterizing the system's law of motion. TSS naturally emerge in many scientific fields, including biology, chemistry, physics, and operations research, and often exhibit complex stochastic dynamics. The existing body of knowledge on TSS comprises of a small number of mathematical models that were introduced throughout the years. Each of these models was tailored for the mathematical modeling of a tandem stochastic system whose behavior is determined by a particular set of rules. Facing the significant complexity inherent to the analysis of general TSS this ad-hoc approach is still considered inevitable. In what follows, we review several types of TSS together with their accompanying mathematical models. Special emphasis is given to the Asymmetric Simple Inclusion Process (ASIP) — a TSS which stands at the heart of this thesis.

3.1 The Tandem Jackson Network

The Tandem Jackson Network (TJN) is one of the first TSS ever to be studied. It was introduced and analyzed by R.R.P. Jackson when he was working for the operational research branch of the London airport [7, 8]. Jackson was inspired by a visit to a factory in which aircraft engines were overhauled. As he explains in the introduction to his paper: *"...Work was carried out on the engines in successive stages, e.g. stripping, detailed examination, repairs, assembly and testing, and thus engines could experience "inter-phase" queueing. It was thought that a mathematical investigation into such a system would be helpful in planning future work and increasing the present efficiency."*. An apology is then quick to come: *"...Unfortunately, to date, the only system of this type which appears to be mathematically tractable is one which is completely random in character, and this is now investigated"*.

In its simplest version, the TJN is a sequential array of n service stations, where external jobs arrive at the leftmost station randomly in time and progress

sequentially from station to station. At each station: (i) arriving jobs queue up in line and await service; (ii) only one job is served at a time, and the service durations are governed by 'exponential clocks'; and (iii) upon completion of service a single job moves on to the next station or, in the case of the rightmost station, out of the system.

In the standard Queueing theory setting, the TJN can be described as a sequential array of Markovian queues [9]. Jobs arrive to the first queue according to a Poisson process with rate λ and are processed, one by one and according to order of arrival, with rate μ_k at station k. Denoting the number of jobs present in the k^{th} station ($k = 1, \cdots, n$) by X_k, the TJN's dynamics can be schematically summarized as follows: (i) first station ($k = 1$):

$$X_1, X_2, \cdots \xrightarrow{\lambda} X_1 + 1, X_2, \cdots ; \tag{1}$$

(ii) interior stations (when $X_k > 0$, $1 < k \leq n-1$):

$$\cdots, X_{k-1}, X_k, X_{k+1}, \cdots \xrightarrow{\mu_k} \cdots, X_{k-1}, X_k - 1, X_{k+1} + 1, \cdots ; \tag{2}$$

(iii) last station (when $X_n > 0$):

$$\cdots, X_{n-1}, X_n \xrightarrow{\mu_n} \cdots, X_{n-1}, X_n - 1 \, . \tag{3}$$

When $n = 1$, the TJN is composed of a single service station and, using the notation introduced by Kendall [10], is equivalent to a simple $M/M/1$ queue. The steady state distribution of this queue is given by

$$Pr(X_1 = x_1) = (1 - \rho_1)\rho_1^{x_1} \, , \tag{4}$$

where $\rho_1 = \lambda/\mu_1$ ($x_1 = 0, 1, 2, \cdots$). While this result was already known to Jackson and his contemporaries, its extension to $n > 1$ stations was considered nontrivial. Indeed, when the output from one station is the input of the next, intricate correlations may render the joint probability distribution of the system extremely complex or even completely intractable.

In light of the above, it is quite remarkable that the steady state distribution of a TJN with n service stations is given by

$$Pr(X_1 = x_1, \cdots, X_n = x_n) = \prod_{k=1}^{n} (1 - \rho_k)\rho_k^{x_k} \, , \tag{5}$$

where $\rho_k = \lambda/\mu_k$ ($x_k = 0, 1, 2, \cdots$). Equation (5) is known as Jackson's theorem, it asserts that stations in the TJN behave *as if* they were a collection of n separate $M/M/1$ queues that are statistically independent of each other. The great simplicity and elegance of this result is best appreciated when compared to other TSS to be described hereinafter.

The TJN is perhaps the simplest queue network imaginable and it is only natural to ask what happens when the model is extended to take into account networks of arbitrary topology and general job routing schemes. The answer to

this question was provided by J.R. Jackson who showed that the steady state distribution of these systems is still given by a product form [11, 12]. The shared surname with R.R.P. has however caused confusion among many (including the author signed above) and this is a good opportunity to set things straight: the TJN and the product form associated with it are known as Jackson's theorem (R.R.P.), Jackson's networks are due to Jackson (J.R.) [13].

It is important to note that Jackson's networks are not limited to constant arrival and service rates. Both R.R.P. and J.R. Jackson recognized the fact that real production systems may, as the amount of work-in-process grows, reduce the rate at which new work is injected or increase the rate at which processing takes place. Indeed, the original product form result of R.R.P Jackson also holds true for sequential networks of $M/M/c$ queues [8]. In these systems, the service rate first rises linearly with the number of jobs in the queue and then flattens at a constant value. This happens because the k^{th} queue is now equipped with $c_k \geq 1$ identical servers — each with service rate μ_k. These servers can simultaneously serve up to c_k jobs in parallel and the effective service rate is thus given by $min(X_k, c_k) \cdot \mu_k$. This somewhat particular scenario may be extended to networks of general topology in which service rates depend almost arbitrarily upon the number of jobs present in the queue of interest and the arrival rate depends almost arbitrarily upon the total number of jobs in the system [12].

Jackson networks were the first significant development in queueing networks theory [14]. Devised in the early sixties of the twentieth century, Jackson networks were further advanced upon to form a theoretical foundation suitable for the analysis of the then emerging packet-switched networks (e.g., the ARPANET) which have by now fully evolved into today's World Wide Web [15]. Since then, the work of J.R. Jackson has found many other applications and has recently received wide-spread recognition when it was re-printed in a special issue dedicated to the 'Ten Most Influential Titles of Management Science's First Fifty Years' [16].

Jackson networks were rediscovered in 1970 by Frank Spitzer when he considered various systems of interacting particles [17]. Mediated only through the dependence of the service rate on the number of jobs present in the *same queue*, interactions in Jackson networks are zero-ranged, i.e., they do not allow jobs at two different queues (even if adjacent) to "feel" each other. In the statistical-physics community, Jackson networks are hence better known as the "Zero-Range Process" [18, 19]. Other types of interaction were also studied by Spitzer and the prominent example of "excluded volume interactions" or "exclusion" will now be discussed. .

3.2 The Asymmetric Simple Exclusion Process

The Asymmetric Simple Exclusion Process (ASEP) — a stochastic process taking place on a discrete one-dimensional lattice of n sites — is a paradigmatic model in non-equilibrium statistical physics [20, 21, 22, 23]. Having first appeared in the literature as a model of bio-polymerization [24], it was later introduced to the probability theory and statistical-physics communities by Frank

Spitzer [17], and has now become a default model for stochastic transport with excluded volume interactions. Over the years, the ASEP and its variants were used to study a wide range of physical phenomena: transport across membranes [25], transport of macromolecules through thin vessels [26], hopping conductivity in solid electrolytes [27], reptation of a polymer in a gel [28], traffic flow [29], gene translation [30, 6], surface growth [31, 32], sequence alignment [33], molecular motors [34] and the directed motion of tracer particles in the presence of dynamical backgrounds [35, 36, 37, 38].

In the ASEP, particles flow randomly in time, into the leftmost site of a one-dimensional lattice and propagate unidirectionally (to the right) along the lattice. Particles move from a site to its right-neighboring site randomly in time — the hopping restricted by the exclusion principle which allows sites to be occupied by no more than one particle at a time. At the rightmost site, particles exit the system randomly in time. The exclusion principle causes jamming throughout the lattice, and renders the ASEP's dynamics highly nontrivial.

The translation between the Queueing-Theory setting of the TJN and the statistical-physics setting of the ASEP is straightforward: 'jobs' are 'particles' and 'service stations' are 'sites'. The random inflow into the leftmost site in the ASEP, the random instants of hopping from site to site, and the random outflow from the rightmost site are all governed by independent Poisson processes. Denoting an occupied site by $\bullet$ and an empty site by $\circ$, the ASEP's dynamics can be schematically summarized as follows: (i) first site ($k = 1$):

$$\circ, \cdots \xrightarrow{\lambda} \bullet, \cdots ; \tag{6}$$

(ii) interior sites ($1 < k \leq n-1$):

$$\cdots, \bullet, \circ, \cdots \xrightarrow{\mu_k} \cdots, \circ, \bullet, \cdots ; \tag{7}$$

(iii) last site ($k = n$):

$$\cdots, \bullet \xrightarrow{\mu_n} \cdots, \circ . \tag{8}$$

Recalling that the capacity of a queue is the maximum number of jobs allowed in it (including those in service); One can think of the ASEP as a TJN of Markovian queues with single job capacity. When this number is reached, further arrivals to the first site are turned away and blocking occurs in interior sites.

In contrast to the steady state distribution of the TJN, that of the ASEP does not obey a product form, i.e., particle occupancies in distinct sites are statistically dependent. It is nevertheless interesting to note that exact expressions for the steady state distribution can sometimes be put in the form of a matrix-product. To this end, the statistical weight of each of the 2^n possible configurations of an ASEP lattice is constructed as a product of matrices, one for each site, chosen according to the state of the site (occupied or empty). The probability to observe the lattice in an particular configuration can then be obtained following proper normalization. A matrix-product solution for the steady state distribution of the ASEP was first derived in [39]. Matrix-product forms are reviewed in [23].

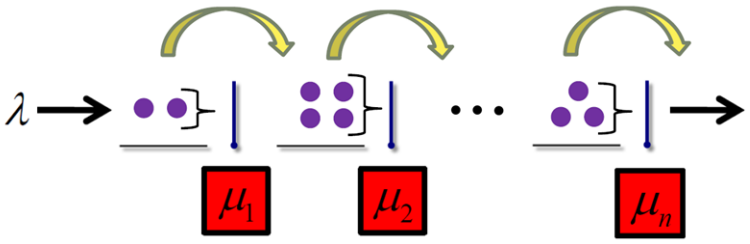

Figure 1: An Illustration of the ASIP.

3.3 The Asymmetric Simple Inclusion Process

Exclusion is central to the ASEP and while this principle is often suitable for the description of some physical systems, it is not suitable for the description of others. Altering the ASEP such that arbitrarily many particles are allowed to simultaneously occupy any given site, one ends up with two different models: the tandem Jackson network (discussed in Subsection 3.1 above), and the Asymmetric Simple Inclusion Process (ASIP), a model which was introduced and analyzed in [1, 2, 3, 5] and is the subject of this thesis. The ASIP is similar to the ASEP — albeit replacing the exclusion principle by an inclusion principle. In both processes, random events cause particles to propagate unidirectionally along a one-dimensional lattice. In the ASEP particles are subject to *exclusion* interactions that keep them singled apart, whereas in the ASIP particles are subject to *inclusion* interactions that coalesce them into inseparable clusters.

The formulation of the ASIP is as follows. Consider a one-dimensional lattice of n sites indexed $k = 1, \cdots, n$. Each site is followed by a gate — labeled by the site's index — which controls the site's outflow. Particles arrive at the first site ($k = 1$) following a Poisson process with rate λ, the openings of gate k are timed according to a Poisson process with rate μ_k ($k = 1, \cdots, n$), and the $n+1$ Poisson processes are mutually independent. A key feature of the ASIP is its 'batch service' property: at an opening of gate k all particles present at site k transit simultaneously, and in one batch (one cluster), to site $k+1$ — thus joining particles that may already be present at site $k+1$ ($k = 1, \cdots, n-1$). At an opening of the last gate ($k = n$) all particles present at site n exit the lattice simultaneously. Denoting the number of particles present in site k ($k = 1, \cdots, n$) by X_k, the ASIP's dynamics can be schematically summarized as follows: (i) first site ($k = 1$):

$$X_1, X_2, \cdots \xrightarrow{\lambda} X_1 + 1, X_2, \cdots ; \tag{9}$$

(ii) interior sites ($1 < k \leq n-1$):

$$\cdots, X_{k-1}, X_k, X_{k+1}, \cdots \xrightarrow{\mu_k} \cdots, X_{k-1}, 0, X_{k+1} + X_k, \cdots ; \tag{10}$$

(iii) last site ($k = n$):

$$\cdots, X_{n-1}, X_n \xrightarrow{\mu_n} \cdots, X_{n-1}, 0\,. \tag{11}$$

	$c_{site}=1$	$c_{site}=\infty$
$c_{gate}=1$	ASEP	TJN
$c_{gate}=\infty$	ASEP	ASIP

Table 1: Capacity classification of the TJN, ASEP, and ASIP models.

The ASIP is further illustrated in Figure 1.

Interestingly, all three models — TJN, ASEP, and ASIP — share the aforementioned sites-gates lattice structure. To pinpoint the difference between the models consider the two following characteristic capacities: (i) *site capacity* c_{site} — the maximal number of particles that can simultaneously occupy a given site, and (ii) *gate capacity* c_{gate} — the maximal number of particles that are simultaneously transferred through a given gate when it opens. In each of the above-mentioned models particles propagate according to the following rule: at an opening of gate k, $min(X_k, c_{site} - X_{k+1}, c_{gate})$ particles transit simultaneously from site k to site $k+1$ — thus joining particles that may already be present at site k ($k = 1, 2, \cdots, n-1$). At an opening of the last gate ($k = n$), $min(X_k, c_{gate})$ particles exit the lattice simultaneously.

In the ASEP the site capacity is $c_{site} = 1$ and the gate capacity can be any positive integer $1 \leq c_{gate} \leq \infty$. In the TJN the site capacity is $c_{site} = \infty$ and the gate capacity is $c_{gate} = 1$. In the ASIP the site capacity is $c_{site} = \infty$ and the gate capacity is $c_{gate} = \infty$. The capacity classification is summarized in Table 1 — from which it is evident that the ASIP is, in effect, a 'missing puzzle piece' connecting together the well established and the well studied ASEP and TJN models.

From a queueing theory perspective, the ASIP is a sequential array of Markovian queues with unbounded capacity and unlimited batch service [40, 41]: all particles present at a given service station are served collectively (and thus move together to the next service station or out of the system). The notion of 'batch service' is strongly related to growth-collapse processes. Consider a single service station with batch service. Jobs arrive to the station randomly in time — causing the queue to grow steadily; when service is rendered all jobs are served simultaneously — causing the queue to collapse to zero. Thus, stochastic growth-collapse temporal patterns emerge from the application of batch-service policies [40, 41, 42, 43, 44]. Interestingly, these patterns appear in a variety of complex systems, including: sand-pile models and systems in self-organized criticality [45], stick-slip models of interfacial friction [46], Burridge-Knopoff type models of earthquakes and continental drift [47], stochastic avalanche models [48], stochastic Ornstein-Uhlenbeck capacitors [49] and geometric Langevin equations [50]. The ASIP model is, in effect, a tandem array of growth-collapse processes.

From a statistical physics perspective, the ASIP is a model for unidirectional transport with coagulation. Such reaction-diffusion models have been extensively studied since the pioneering work of Smoluchowski [51]. Yet still, unresolved issues and intriguing new facets cause them to raise interest even to-

day [52, 53]. Two of the simplest models of this kind are the coalescence-diffusion model

$$\begin{aligned} \cdots \bullet \bullet \cdots &\xrightarrow{1} \cdots \circ \bullet \cdots \ , \\ \cdots \bullet \circ \cdots &\xrightarrow{1} \cdots \circ \bullet \cdots \ , \end{aligned} \tag{12}$$

where $\bullet$ represents an occupied site and $\circ$ represents an empty site, and the aggregation-diffusion model

$$\begin{aligned} \cdots A_l A_{l'} \cdots &\xrightarrow{1} \cdots 0 A_{l+l'} \cdots \ , \\ \cdots A_l 0 \cdots &\xrightarrow{1} \cdots 0 A_l \cdots \ , \end{aligned} \tag{13}$$

where A_l represents a site occupied by $l > 0$ particles, and 0 represents an empty site [54].

The studies dedicated to the models described in Eqs. (12) and (13) were, by and large, carried out in a one-dimensional ring topology. Under these conditions many statistical properties can be calculated exactly using the empty-interval method [54, 55], which we shall address in chapter 8. The ASIP, with homogeneous unit rates $\{\mu_1 = \cdots = \mu_n = 1\}$, can be viewed as a generalization of aggregation-diffusion models to an open system. Indeed, the bulk ASIP dynamics of Eq. (10) is identical to the dynamics of Eq. (13). Similarly, when one disregards the number of particles occupying each site (X_k) and focuses only on whether sites are occupied or not ($X_k > 0$ or $X_k = 0$), the ASIP dynamics turns into an open-boundary version of Eq. (12). In the following chapter, we combine probabilistic and Monte-Carlo analysis to explore the ASIP and demonstrate that it is a true showcase of statistical complexity.

4 A Showcase of Complexity

In this chapter we motivate the study of the ASIP as a showcase of complexity. We use Monte-Carlo simulations to unveil a rich assortment of statistical behaviors which manifest the ASIP's intrinsic complexity. We focus on the steady state of large ($n \gg 1$) homogeneous ASIPs that are characterized by identical service rates: $\mu_1 = \cdots = \mu_n$. In Chapter 5, we show that this particular subclass of ASIPs is optimal with respect to various measures of efficiency and is hence of special importance. Here, we are mainly interested in demonstrating that complex behavior is observed even in the simplest ASIP systems imaginable and further set $\lambda = \mu_1 = \cdots = \mu_n = 1$.

In what follows we denote by $X_k(t)$ the number of particles present in site k at time t, and set $\mathbf{X}(t) = (X_1(t), \cdots, X_n(t))$. The random vector $\mathbf{X}(t)$ represents the ASIP's occupancy at time t ($t \geq 0$). In Chapter 5, we show that the stochastic processes $(\mathbf{X}(t))_{t \geq 0}$ is asymptotically *stationary,* and that it converges in law (as $t \to \infty$) to a stochastic limit $\mathbf{X} = (X_1, \cdots, X_n)$. The random variable X_k represents the number of particles present at site k at steady state, and the random vector $\mathbf{X}$ represents the ASIP's occupancy at steady state. Henceforth, given a vector $\mathbf{v} = (v_1, ..., v_n)$ we denote by $|\mathbf{v}| = v_1 + ... + v_n$ its sum of coordinates, and by $\#(\mathbf{v})$ the number of its non-zero coordinates.

In many systems results obtained by *mean field analysis* provide a fair approximation to the systems' steady state behavior. This holds for the ASEP and TJN, but does *not* hold for the ASIP. In Chapter 5, we show that the ASIP's mean occupancy at steady state is given by $\langle X_k \rangle = \lambda / \mu_k$ ($k = 1, \cdots, n$). Thus, for the homogeneous ASIPs discussed in this chapter $\langle X_k \rangle = 1$. On the other hand, Monte-Carlo analysis depicted in top panel of Figure 2 asserts that the following *power-law asymptotics* hold ($k \gg 1$):

$$\begin{cases} \Pr(X_k > 0) \approx k^{-1/2} , \\ \\ \langle X_k \mid X_k > 0 \rangle \approx k^{1/2} , \\ \\ \sigma(X_k) / \langle X_k \rangle \approx k^{1/4} , \end{cases} \tag{14}$$

where $\sigma(X_k)$ denotes the standard deviation of the random variable X_k. Namely, at steady state: (i) the probability that site k is occupied decreases like $k^{-1/2}$; (ii) the conditional mean number of particles occupying site k, given that the site is occupied, increases like $k^{1/2}$ [57]; (iii) the standard deviation of the number of particles occupying site k, measured with respect to the mean number of particles occupying site k, increases like $k^{1/4}$. The power-law asymptotics of Eq. (14) imply that 'downstream sites' ($k \gg 1$) are rarely occupied, but when they are — they are occupied by a large number of particles (in comparison to the mean occupancy $\langle X_k \rangle = 1$). This 'all-or-none' type of steady-state behavior results in large occupancy fluctuations of downstream sites, and hence renders the mean field approach rather limited for the ASIP.

The power-law asymptotics of Eq. (14) are further induced to the ASIP's density of occupied sites: $D_n = \frac{1}{n} \# (\mathbf{X})$. A Monte-Carlo analysis depicted in

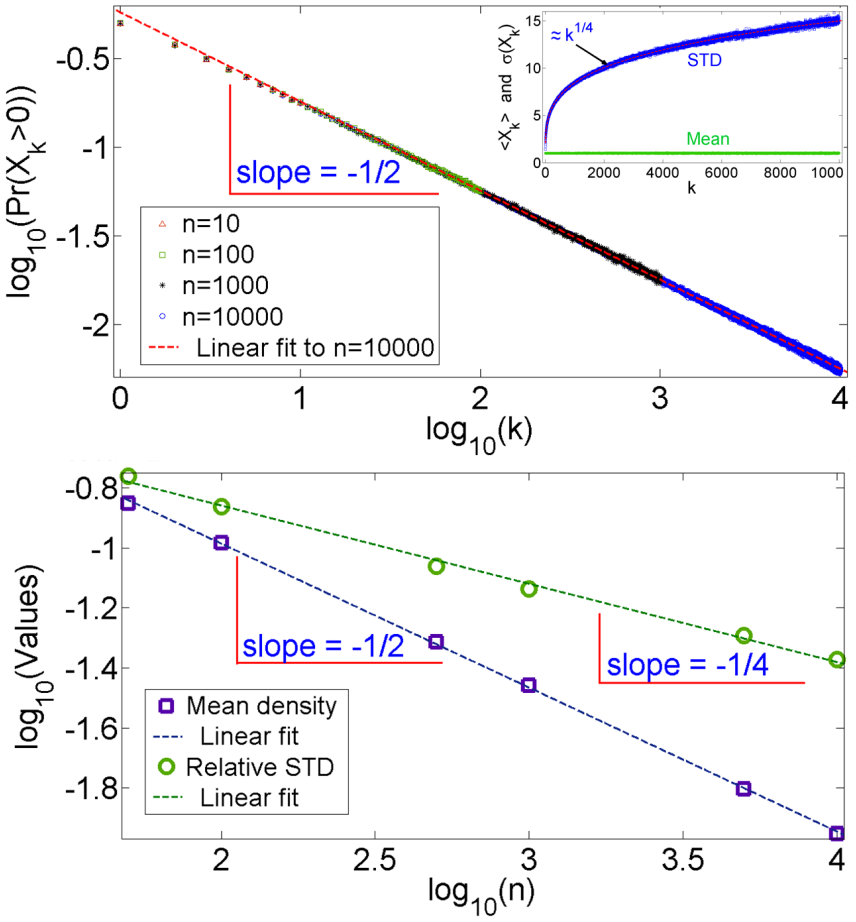

Figure 2: Top panel main: The probability that site k is occupied, as a function of the index k, on a log-log plot; the linear fit implies a power law decay with exponent $-1/2$. Top panel inset: The mean and the standard deviation (STD) of the occupancy of site k, as a function of the index k. Each site is occupied, on average, by a single particle, yet the fluctuations around the mean grow like $k^{1/4}$, and are hence typically much larger than the mean occupancy itself. Bottom panel: The mean and the relative STD of the density of occupied sites, as a function of the lattice size n, on a log-log plot; the linear fits imply power law decays with respective exponents $-1/2$ and $-1/4$ — which, in turn, manifests condensation.

the bottom panel of Figure 2 asserts that the following *power-law asymptotics* hold ($n \gg 1$):

$$\begin{cases} \langle D_n \rangle \approx n^{-1/2} \ , \\ \\ \sigma\left(D_n\right) / \langle D_n \rangle \approx n^{-1/4} \ , \end{cases} \tag{15}$$

where $\sigma\left(D_n\right)$ denotes the standard deviation of the density D_n. Namely, at steady state the particles occupying the lattice sites *condense* to a vanishingly small fraction D_n of sites: (i) the mean density of occupied sites decreases to zero like $n^{-1/2}$; (ii) the standard deviation of the density of occupied sites, measured with respect to the mean density of occupied sites, decreases like $n^{-1/4}$. This *power-law condensation* of particles is yet another manifestation of the 'all-or-none' steady-state behavior noted above.

We now turn to explore four key observables Θ_n of the ASIP at steady state: *Load*, *draining time*, *inter-exit time*, and *coalescence time*. A Monte-Carlo analysis asserts that these four observables (to be defined momentarily) are random variables admitting asymptotic stochastic approximations of the form ($n \gg 1$):

$$\Theta_n \approx a_n \cdot \Theta + b_n \ , \tag{16}$$

where a_n and b_n are deterministic scaling coefficients, and where Θ is a limiting random variable. For each observable the coefficients a_n, b_n, and the limit Θ will be specified hereinafter. We note that the three random times explored below — the draining time, the inter-exit time, and the coalescence time — are, in effect, first passage times of the ASIP [58].

Load. The ASIP's load is the total number of particles present in the lattice: $\Theta_n = |\mathbf{X}|$. For the ASIP's load the coefficients are $a_n = \sqrt{2n}$ and $b_n = n$, and the limit Θ is *Gaussian* with zero mean and unit variance. This result is identical, in form, to the standard Central Limit Theorem (CLT) [59]. However, while the standard CLT setting requires the random variables $\{X_k\}_{k=1}^n$ to be independent and identically distributed, in the ASIP the random variables $\{X_k\}_{k=1}^n$ are neither independent nor identically distributed. In Chapter 5 we establish that the steady state load of an ASIP with rates $(\lambda, \mu_1, \cdots, \mu_n)$ is equal, in law, to the sum of loads of n *independent single-site* ASIPs with respective rates $(\lambda, \mu_1), \cdots, (\lambda, \mu_n)$. Thus, in the case of homogeneous ASIPs, the CLT can be applied to obtain the load asymptotics.

Draining time. Consider the ASIP with no inflow ($\lambda = 0$), and assume that at time $t = 0$ the ASIP's occupancy is given by the steady state vector: $\mathbf{X}(0) = \mathbf{X}$. The ASIP's draining time is the time elapsing till the lattice is clear of particles: $\Theta_n = \inf\{t \geq 0 \mid |\mathbf{X}(t)| = 0\}$. In other words, the ASIP's draining time is the random time required for 'draining out' an ASIP at steady state, after having blocked the inflow of new-coming particles. For the ASIP's draining time the coefficients are $a_n = \sqrt{n}$ and $b_n = n$, and the limit Θ is *Gaussian* with zero mean and unit variance. In effect, Monte-Carlo analysis illustrated in Figure 3 asserts that the draining time Θ_n is approximately *Gamma* with mean n and variance n. In turn, this Gamma approximation implies that the draining time Θ_n is equal, in law, to the sum of n *independent Exponential* random variables

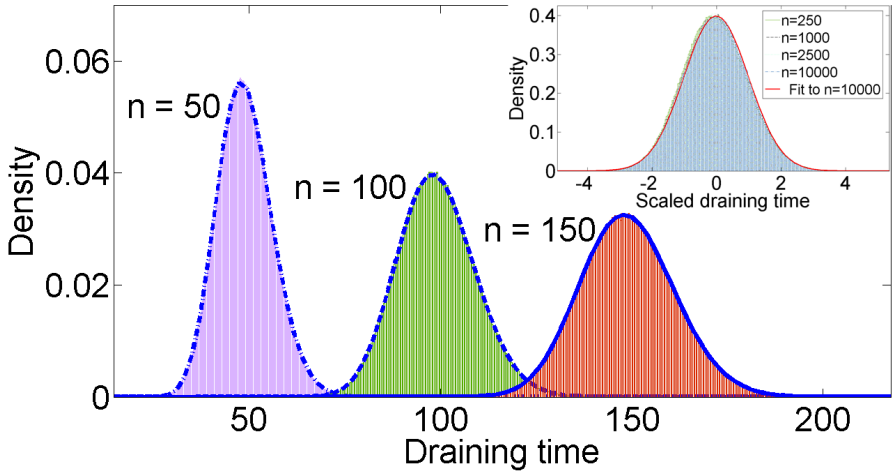

Figure 3: Main: Gamma approximation of the draining time; bars represent simulated histograms, and dashed lines represent Gamma density fits. The mean and the variance of the Gamma distribution, for $n = [50, 100, 150]$, are given respectively by $[49, 99, 149]$ and $[51.1, 101.3, 151.2]$. Inset: The Gaussian limit of the scaled draining time.

with unit mean [59]. Thus, the CLT applies to the draining-time asymptotics as well.

Inter-exit time. The openings of the last ASIP gate are governed by a Poisson process with unit rate ($\mu_n = 1$), and when the last gate opens all the particles present in the last site exit the lattice. Equation (14) asserts that the steady-state probability that the last site is non-empty is given by $\Pr(X_n > 0) \approx n^{-1/2}$. Consequently, not every opening of the last gate indeed results in an exit of particles from the lattice. The ASIP's inter-exit time — for an ASIP in steady state — is defined as the time elapsing between two consecutive time epochs at which particles exit the lattice. For the ASIP's inter-exit time the coefficients are $a_n = \sqrt{\pi n}$ and $b_n = 0$, and the limit Θ is *Rayleigh* with unit mean and probability tail

$$\Pr(\Theta > t) = \exp\left(-\pi t^2/4\right) \tag{17}$$

($t > 0$). In effect, the Monte-Carlo analysis illustrated in Figure 4 shows that the Rayleigh approximation well captures the data even for relatively small n.

The hazard rate $h_T(t)$ ($t > 0$) of a random time T is defined as the limit $h_T(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr(T \leq t + \delta \mid T > t)$ [60]. Namely, given that the random time T did not realize during the time interval $[0, t]$, the realization rate of random time T immediately after time t is $h_T(t)$. On the one hand, the ASIP's inter-arrival time is Exponential — which is the unique random time characterized by a *constant* hazard rate. On the other hand, the ASIP's inter-exit time is approximately Rayleigh — which is the unique random time characterized by

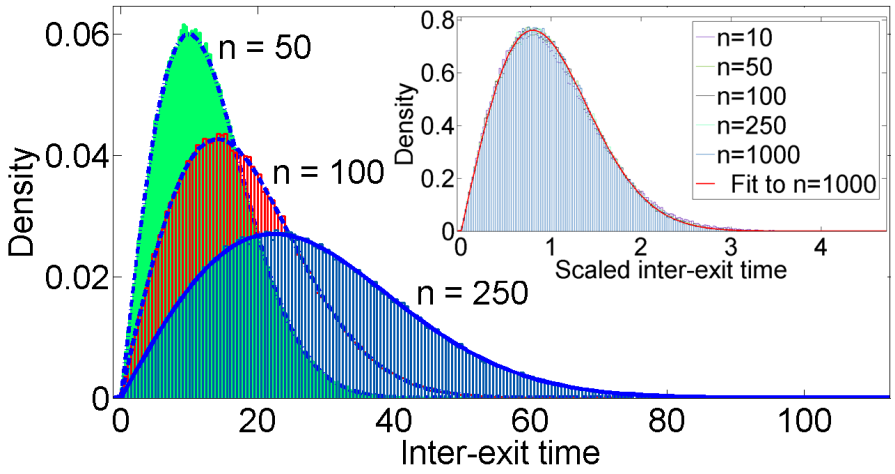

Figure 4: Main: Rayleigh approximation of the inter-exit time; bars represent simulated histograms, and dashed lines represent Rayleigh density fits. The mean and variance of the Rayleigh distribution, for $n = [50, 100, 250]$, are given respectively by $[12.6, 17.8, 28.1]$ and $[43.3, 86.2, 215.4]$. Inset: The Rayleigh limit of the scaled inter-exit time.

a *linear* hazard rate. Namely, the inter-arrival time is memory-less (due to its constant hazard rate), whereas the inter-exit time is aging linearly: $h_{\Theta}(t) = (\pi/2) \cdot t$. Thus, in the transition from the ASIP's inflow to the ASIP's outflow an *aging effect* emerges.

Coalescence time. Consider a *circular* ASIP in which the output from the last site is the input of the first site, and assume that at time $t = 0$ all sites are occupied: $\#(\mathbf{X}(0)) = n$. As time progresses, gates open and particles coalesce into larger and larger particle-clusters. Eventually, all particles will coalesce to a single 'super cluster'. The ASIP's coalescence time is the time elapsing till all particles coalesce together and form the 'super cluster': $\Theta_n = \inf\{t \geq 0 \mid \#(\mathbf{X}(t)) = 1\}$. For the ASIP's coalescence time the coefficients are $a_n = n^2/6$ and $b_n = 0$, and the limit Θ is *Inverse Gaussian* with unit mean and probability density function

$$\frac{d}{dt}\Pr(\Theta \leq t) = \frac{1}{\sqrt{2\pi\nu}} \cdot t^{-3/2} \exp\left(-\frac{(t-1)^2}{2\nu t}\right) \tag{18}$$

($t > 0$; $\nu = 2/5$). In effect, the Monte-Carlo analysis illustrated in Figure 5 shows that the Inverse-Gaussian approximation well captures the data even for relatively small n.

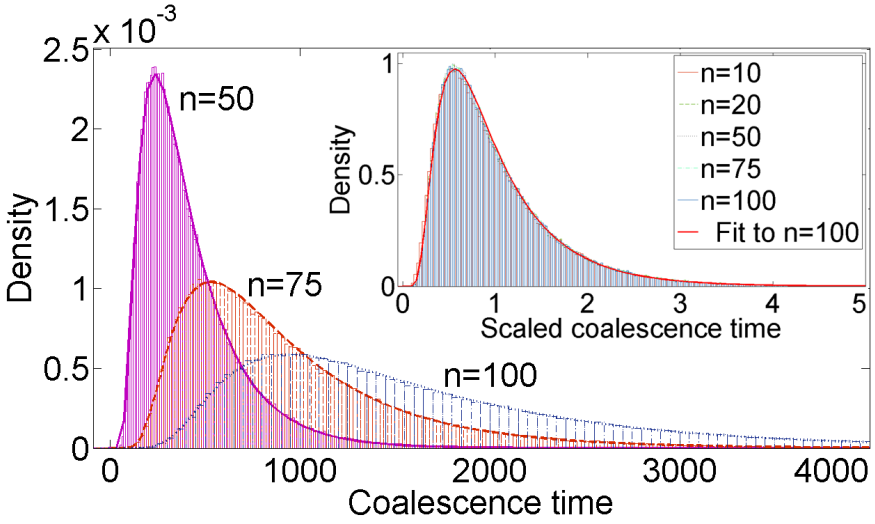

Figure 5: Main: Inverse-Gaussian approximation of the coalescence time; bars represent simulated histograms, and dashed lines represent Inverse-Gaussian density fits. The mean and the variance, for $n = [50, 75, 100]$, are given respectively by $[416, 937, 1669]$ and $[69984, 354334, 1126670]$. Inset: The Inverse-Gaussian limit of the scaled coalescence time.

	a_n	b_n	Θ	$\langle\Theta\rangle$	$\sigma^2(\Theta)$
Load	$\sqrt{2n}$	n	Gaussian	0	1
Draining time	$\sqrt{n}$	n	Gaussian	0	1
Inter-exit time	$\sqrt{\pi n}$	0	Rayleigh	1	$\frac{4-\pi}{\pi}$
Coalescence time	$n^2/6$	0	Inv. Gauss.	1	$2/5$

Table 2: Asymptomatic stochastic approximations of the ASIP observables: Load, draining time, inter-exit time, and coalescence time.

In this chapter we have shown that the statistical behavior of the ASIP is rich — ranging from various 'mild' and 'intermediate' forms of randomness (displayed by the load and first passage times), to 'wild' forms of randomness (displayed by the occupancies and by the density of occupied sites) [61]. The four asymptotic stochastic approximation results are summarized in Table 2 above. In the following chapters we will study the ASIP with analytical tools and rigorously establish a comprehensive characterization of its steady state statistics.

5 First Steps

In this chapter we make first steps in the analysis of the ASIP model. Our focus is set on the stochastic dynamics and the stationary statistics of the ASIP's occupancy-vector: the n-dimensional vector counting the number of particles present in each lattice site (at any given time). We derive evolution equations and steady-state equations for the mean and for the Probability Generating Function (PGF) of the ASIP's occupancy-vector. Explicit steady state solutions are obtained for the mean. Explicit steady-state solutions are also obtained for the PGF of small ASIP systems ($n = 1, 2, 3$), and a computational scheme for solving the steady-state PGF equations for ASIP systems of arbitrary size is presented. We show that the steady-state PGF solutions 'explode' in complexity as the lattice-size increases — thus rendering the ASIP's occupancy vector analytically intractable for large n.

The ASIP's load is the total number of particles present in the lattice. In comparison to the ASIP's occupancy-vector, analytical tractability of the ASIP's load is much more simple. Indeed, we obtain closed-form results for the mean, variance, and PGF of the ASIP's load in steady-state. Interestingly, the load's PGF admits a product form representation — which, in turn, implies a surprising stochastic decomposition structure. Moreover, with the explicit steady-state load results at hand, we further study load-optimization in steady-state, seeking system-parameters — namely, the rates of the underlying exponential clocks — that optimize the ASIP in various aspects. Our analysis concludes that optimality is attained by *homogenous* ASIP systems in which the underlying 'exponential clocks' all have the same rate.

The reminder of the chapter is organized as follows. In Sections 5.1–5.3 we describe the traversal time, derive the ASIP's Markovian 'law of motion', and present a Monte Carlo algorithm for the simulation of its stochastic evolution. The mean analysis and the PGF analysis of the ASIP are carried out, respectively, in Sections 5.4 and 5.5. The solution of the ASIP's PGF in steady state is discussed in Section 5.6. The ASIP's load and load-optimization are analyzed, respectively, in Sections 5.7 and 5.8.

5.1 Traversal Time

Consider the random time T it takes a particle to traverse the system — henceforth termed the ASIP's '*traversal time*'. That is, T is the time elapsing from the instant a particle arrives at the first site, till the instant it leaves the system. Due to the memory-less property of the exponential distribution, the time elapsing from the arrival of a particle to site k (at an arbitrary time epoch), till the first opening of gate k thereafter, is exponentially distributed with mean $1/\mu_k$. A particle arriving to the system would thus wait an exponentially-distributed random time (with mean $1/\mu_1$) till moving from the first site to the second site, then wait an exponentially-distributed random time (with mean $1/\mu_2$) till moving from the second site to the third site, and so forth. Since the gate-openings are governed by independent Poisson processes we conclude that the traversal

time T admits the following stochastic representation

$$T = \triangle_1 + \cdots + \triangle_n \ , \tag{19}$$

where $\{\triangle_1, \cdots, \triangle_n\}$ is a sequence of independent and exponentially-distributed random times with corresponding means $\{1/\mu_1, \cdots, 1/\mu_n\}$. Consequently, the mean and the variance of the traversal time T are given, respectively, by

$$\mathbf{E}\left[T\right] = \frac{1}{\mu_1} + \cdots + \frac{1}{\mu_n} \tag{20}$$

and

$$\mathbf{Var}\left[T\right] = \frac{1}{\mu_1^2} + \cdots + \frac{1}{\mu_n^2} \ . \tag{21}$$

Henceforth, we shall use the shorthand notation $\mu = \mu_1 + \cdots + \mu_n$ for the system's cumulative 'service rate'.

5.2 Markovian dynamics

Let $X_k\left(t\right)$ denote the number of particles present in the k^{th} site ($k = 1, \cdots, n$) at time t ($t \geq 0$), and set $\mathbf{X}\left(t\right) = \left(X_1\left(t\right), \cdots, X_n\left(t\right)\right)$. The vector $\mathbf{X}\left(t\right)$ represents the system's occupancy at time t. Observe the system at times t and $t' = t + \Delta$ (for small Δ) and use the shorthand notation $\mathbf{X} = \mathbf{X}\left(t\right)$ and $\mathbf{X}' = \mathbf{X}\left(t'\right)$. The stochastic connection between the random vectors $\mathbf{X}$ and $\mathbf{X}'$ — characterizing the Markovian 'law of motion' of the stochastic process $\left(\mathbf{X}\left(t\right)\right)_{t \geq 0}$ — is given by:

$$\begin{aligned}
&\left(X_1', \cdots, X_n'\right) \\
&= \begin{cases}
\left(X_1, X_2, X_3, \cdots, X_{n-1}, X_n\right) & \textbf{w.p. } 1 - \left(\lambda + \mu\right)\Delta + o(\Delta) \ , \\
\left(X_1 + 1, X_2, X_3, \cdots, X_{n-1}, X_n\right) & \textbf{w.p. } \lambda\Delta + o(\Delta) \ , \\
\left(0, X_1 + X_2, X_3, \cdots, X_{n-1}, X_n\right) & \textbf{w.p. } \mu_1\Delta + o(\Delta) \ , \\
\left(X_1, 0, X_2 + X_3, \cdots, X_{n-1}, X_n\right) & \textbf{w.p. } \mu_2\Delta + o(\Delta) \ , \\
\vdots & \vdots \\
\left(X_1, X_2, X_3, \cdots, 0, X_{n-1} + X_n\right) & \textbf{w.p. } \mu_{n-1}\Delta + o(\Delta) \ , \\
\left(X_1, X_2, X_3, \cdots, X_{n-1}, 0\right) & \textbf{w.p. } \mu_n\Delta + o(\Delta) \ .
\end{cases}
\end{aligned} \tag{22}$$

Eq. (22) follows from considering the totality of events that may take place within the time interval $(t, t']$. There are $n + 1$ such events, and we label them according to the Poisson processes inducing them: (0) the arrival of a particle to the first site — occurring with probability $\lambda\Delta + o(\Delta)$ — in which case $X_1 \mapsto X_1' = X_1 + 1$; (1) opening of the first gate — occurring with probability $\mu_1\Delta +$

$o(\Delta)$ — in which case $X_1 \mapsto X_1' = 0$ and $X_2 \mapsto X_2' = X_1 + X_2$; (2) opening of the second gate — occurring with probability $\mu_2\Delta + o(\Delta)$ — in which case $X_2 \mapsto X_2' = 0$ and $X_3 \mapsto X_3' = X_2 + X_3$; $\cdots$; $(n-1)$ opening of the gate before last — occurring with probability $\mu_{n-1}\Delta + o(\Delta)$ — in which case $X_{n-1} \mapsto X_{n-1}' = 0$ and $X_n \mapsto X_n' = X_{n-1} + X_n$; (n) opening of the last gate — occurring with probability $\mu_n\Delta + o(\Delta)$ — in which case $X_n \mapsto X_n' = 0$. The first line on the right-hand-side of Eq. (22) represents the scenario in which no event takes place — which occurs with the complementary probability $1 - (\lambda + \mu)\,\Delta + o(\Delta)$.

5.3 Monte Carlo Simulation

The ASIP's random trajectory $(\mathbf{X}(t))_{t\geq 0}$ changes discretely rather than continuously. Indeed, between the underlying 'Poissonian events' — arrival of a particle to the system, or an opening of one of the n gates — the ASIP's trajectory does not change. Consider now the ASIP's trajectory at the instants it changes (i.e., arrival of a particle or an opening of a gate). Let $Y_k(s)$ denote the number of particles present in the k^{th} site $(k = 1, \cdots, n)$ immediately after the s^{th} Poissonian event took place $(s = 1, 2, \cdots)$, and set $\mathbf{Y}(s) = (Y_1(s), \cdots, Y_n(s))$. Observe the system at two consecutive Poissonian events, s and $s' = s + 1$, and use the shorthand notation $\mathbf{Y} = \mathbf{Y}(s)$ and $\mathbf{Y}' = \mathbf{Y}(s')$. The properties of the exponential distribution imply that [59]:

1. The time elapsing between two consecutive Poissonian events s and $s' = s + 1$ is exponentially distributed with mean $1/(\lambda + \mu)$.

2. The stochastic connection between the random vectors $\mathbf{Y}$ and $\mathbf{Y}'$ — characterizing the Markovian 'law of motion' of the stochastic process $(\mathbf{Y}(s))_{s=1}^{\infty}$ — is given by

$$(Y_1', \cdots, Y_n') = \begin{cases} (Y_1 + 1, Y_2, Y_3, \cdots, Y_{n-1}, Y_n) & \textbf{w.p. } \lambda/(\lambda+\mu)\ , \\ (0, Y_1 + Y_2, Y_3, \cdots, Y_{n-1}, Y_n) & \textbf{w.p. } \mu_1/(\lambda+\mu)\ , \\ (Y_1, 0, Y_2 + Y_3, \cdots, Y_{n-1}, Y_n) & \textbf{w.p. } \mu_2/(\lambda+\mu)\ , \\ \vdots & \vdots \\ (Y_1, Y_2, Y_3, \cdots, 0, Y_{n-1} + Y_n) & \textbf{w.p. } \mu_{n-1}/(\lambda+\mu)\ , \\ (Y_1, Y_2, Y_3, \cdots, Y_{n-1}, 0) & \textbf{w.p. } \mu_n/(\lambda+\mu)\ . \end{cases} \tag{23}$$

3. The time elapsing between the two consecutive Poissonian events s and $s' = s + 1$, and the change $\mathbf{Y} \mapsto \mathbf{Y}'$, are mutually independent.

Eq. (23) follows from considering the totality of events that lead to a change $\mathbf{Y} \mapsto \mathbf{Y}'$. There are $n + 1$ such events, and we label them according to the

Poisson processes inducing them: (0) the arrival of a particle to the first site — occurring with probability $\lambda/(\lambda+\mu)$ — in which case $Y_1 \mapsto Y_1' = Y_1 + 1$; (1) opening of the first gate — occurring with probability $\mu_1/(\lambda+\mu)$ — in which case $Y_1 \mapsto Y_1' = 0$ and $Y_2 \mapsto Y_2' = Y_1 + Y_2$; (2) opening of the second gate — occurring with probability $\mu_2/(\lambda+\mu)$ — in which case $Y_2 \mapsto Y_2' = 0$ and $Y_3 \mapsto Y_3' = Y_2 + Y_3$; $\cdots$; $(n-1)$ opening of the gate before last — occurring with probability $\mu_{n-1}/(\lambda+\mu)$ — in which case $Y_{n-1} \mapsto Y_{n-1}' = 0$ and $Y_n \mapsto Y_n' = Y_{n-1} + Y_n$; (n) opening of the last gate — occurring with probability $\mu_n/(\lambda+\mu)$ — in which case $Y_n \mapsto Y_n' = 0$. Properties 1,2 and 3 establish, in effect, a simple and straightforward Monte Carlo algorithm for the simulation of the ASIP's random trajectory.

5.4 Mean Dynamics and Mean Analysis

In this section we study the dynamics of the *means* of the random vectors $\mathbf{X}(t)$ $(t \geq 0)$ and $\mathbf{Y}(s)$ $(s = 1, 2, \cdots)$. Throughout the section we use the shorthand vector notation $\boldsymbol{\lambda} = (\lambda, 0, \cdots, 0)^T$, and the shorthand matrix notation

$$\mathbf{M} = \begin{pmatrix} -\mu_1 & & & \cdots & & \\ \mu_1 & -\mu_2 & & \cdots & & \\ & \mu_2 & -\mu_3 & & & \\ & & \ddots & \ddots & & \\ & & & \mu_{n-2} & -\mu_{n-1} & \\ & & & & \mu_{n-1} & -\mu_n \end{pmatrix} \tag{24}$$

(Note: in the above Matrix all blank spaces represent zero entries).

5.4.1 Mean Dynamics of $\mathbf{X}(t)$

Denote the mean of the random vector $\mathbf{X}(t)$ by

$$\mathbf{e_X}(t) = (\mathbf{E}[X_1(t)], \cdots, \mathbf{E}[X_n(t)])^\top \ . \tag{25}$$

Observe the system at times t and $t' = t+\Delta$. Conditioning on $\mathbf{X}(t)$ and utilizing the Markovian dynamics of Eq. (22) yields

$$\mathbf{E}\left[\mathbf{X}(t')\right] = \mathbf{E}\left[\mathbf{E}\left[\mathbf{X}(t')|\ \mathbf{X}(t)\right]\right]$$

$$= \begin{cases} \left(1-(\lambda+\mu)\,\Delta\right)\mathbf{E}\left[\mathbf{X}(t)\right] + \lambda\Delta\mathbf{E}\left[\mathbf{X}(t) + (1,0,\cdots,0)^{\top}\right] \\ + \\ (\mu_1\Delta)\mathbf{E}\left[\mathbf{X}(t) + \left(-X_1(t), X_1(t), 0, \cdots, 0\right)^{\top}\right] \\ + \\ (\mu_2\Delta)\mathbf{E}\left[\mathbf{X}(t) + \left(0, -X_2(t), X_2(t), \cdots, 0\right)^{\top}\right] \\ + \cdots + \\ (\mu_{n-1}\Delta)\mathbf{E}\left[\mathbf{X}(t) + \left(0, \cdots, 0, -X_{n-1}(t), X_{n-1}(t)\right)^{\top}\right] \\ + \\ (\mu_n\Delta)\mathbf{E}\left[\mathbf{X}(t) + \left(0, \cdots, 0, -X_n(t)\right)^{\top}\right] \\ + \\ o(\Delta)\ . \end{cases} \tag{26}$$

Rearranging the terms of Eq. (26), dividing by Δ, and taking $\Delta \to 0$, we conclude that

$$\frac{d\mathbf{e_X}}{dt}(t) = \mathbf{M}\mathbf{e_X} + \boldsymbol{\lambda}\ . \tag{27}$$

Eq. (27) represents the 'mean dynamics' of the random vector $\mathbf{X}(t)$. Namely, it transforms the Markovian dynamics of Eq. (22) to a differential equation that governs the temporal evolution of the mean vector $\mathbf{e_X}(t)$ $(t \geq 0)$. The solution of Eq. (27) can be shown to be given by

$$\mathbf{e_X}(t) = \mathbf{M}^{-1}\left[exp(\mathbf{M}t) - \mathbf{I}\right]\boldsymbol{\lambda}\ . \tag{28}$$

5.4.2 Mean Dynamics of $\mathbf{Y}(s)$

Denote the mean of the random vector $\mathbf{Y}(s)$ by

$$\mathbf{e_Y}(s) = \left(\mathbf{E}\left[Y_1(s)\right], \cdots, \mathbf{E}\left[Y_n(s)\right]\right)^{\top}\ . \tag{29}$$

Observe the system at two consecutive s and $s' = s+1$ Poissonian events. Conditioning on $\mathbf{Y}(s)$ and utilizing the 'law of motion' presented in Eq. (23)

yields

$$\mathbf{E}\left[\mathbf{Y}(s')\right] = \mathbf{E}\left[\mathbf{E}\left[\mathbf{Y}(s')|\mathbf{Y}(s)\right]\right]$$

$$= \begin{cases} \frac{\lambda}{\lambda+\mu}\mathbf{E}\left[\mathbf{Y}(s) + (1,0,\cdots,0)^{\top}\right] \\ + \\ \frac{\mu_1}{\lambda+\mu}\mathbf{E}\left[\mathbf{Y}(s) + (-Y_1(s), Y_1(s), 0, \cdots, 0)^{\top}\right] \\ + \\ \frac{\mu_2}{\lambda+\mu}\mathbf{E}\left[\mathbf{Y}(s) + (0, -Y_2(s), Y_2(s), \cdots, 0)^{\top}\right] \\ + \cdots + \\ \frac{\mu_{n-1}}{\lambda+\mu}\mathbf{E}\left[\mathbf{Y}(s) + (0, \cdots, 0, -Y_{n-1}(s), Y_{n-1}(s))^{\top}\right] \\ + \\ \frac{\mu_n}{\lambda+\mu}\mathbf{E}\left[\mathbf{Y}(s) + (0, \cdots, 0, -Y_n(s))^{\top}\right] \ . \end{cases} \tag{30}$$

Rearranging the terms of Eq. (30) we conclude that

$$\begin{aligned} &(\lambda+\mu)\left(\mathbf{e}_{\mathbf{Y}}(s') - \mathbf{e}_{\mathbf{Y}}(s)\right) \\ &= \mathbf{M}\mathbf{e}_{\mathbf{Y}}(s) + \boldsymbol{\lambda} \ . \end{aligned} \tag{31}$$

Eq. (31) represents the 'mean dynamics' of the random vector $\mathbf{Y}(s)$. That is, it transforms the 'law of motion' of Eq. (23) to a difference equation that governs the temporal evolution of the mean vector $\mathbf{e}_{\mathbf{Y}}(s)$ $(s = 1, 2, \cdots)$. The solution of Eq. (31) can be shown to be given by

$$\mathbf{e}_{\mathbf{Y}}(s) = \mathbf{M}^{-1}\left[(\mathbf{I} + \frac{1}{\lambda+\mu}\mathbf{M})^s - \mathbf{I}\right]\boldsymbol{\lambda} \ . \tag{32}$$

5.4.3 Mean Analysis in Steady State

Consider now the ASIP model in *steady state*. In steady state the stochastic processes $(\mathbf{X}(t))_{t\geq 0}$ and $(\mathbf{Y}(s))_{s=1}^{\infty}$ are *stationary*, and hence their respective means are time-homogeneous: $\mathbf{e}_{\mathbf{X}}(t) \equiv \mathbf{e}_{\mathbf{X}}$ $(t \geq 0)$ and $\mathbf{e}_{\mathbf{Y}}(s) \equiv \mathbf{e}_{\mathbf{Y}}$ $(s = 1, 2, \cdots)$. Substituting the time-homogeneous vectors $\mathbf{e}_{\mathbf{X}}(t) \equiv \mathbf{e}_{\mathbf{X}}$ and $\mathbf{e}_{\mathbf{Y}}(s) \equiv \mathbf{e}_{\mathbf{Y}}$, respectively, into Eqs. (27) and (31) yields the common equation

$$0 = \mathbf{M}\mathbf{e} + \boldsymbol{\lambda} \ . \tag{33}$$

(where $\mathbf{e} = (e_1, \cdots, e_n)$ is the *unknown* vector). Namely, both the mean vectors $\mathbf{e}_{\mathbf{X}}$ and $\mathbf{e}_{\mathbf{Y}}$ are governed by Eq. (33).

A straightforward computation of Eq. (33) yields the steady state solution

$$e_k = \mathbf{E}\left[X_k(t)\right] = \mathbf{E}\left[Y_k(s)\right] = \frac{\lambda}{\mu_k} \tag{34}$$

$(k = 1, \cdots, n)$. Combining Eqs. (20) and (34) together further yields the following steady state formula:

$$\mathbf{E}\left[\sum_{k=1}^{n} X_k(t)\right] = \mathbf{E}\left[\sum_{k=1}^{n} Y_k(s)\right] = \sum_{k=1}^{n}\frac{\lambda}{\mu_k} = \lambda\mathbf{E}\left[T\right] \ . \tag{35}$$

Equation (35) asserts that at steady state the mean number of particles in the system is given by the product $\lambda \mathbf{E}[T]$: the flow rate λ into the system times the mean traversal time $\mathbf{E}[T]$ — the mean sojourn time of an arbitrary particle in the system. Note that although the random variables $\{X_1(t), \cdots, X_n(t)\}$ (and similarly $\{Y_1(s), \cdots, Y_n(s)\}$) are intricately dependent, these dependencies do not affect the mean behavior given by Eq. (35). Equation (35) is the 'ASIP version' of the well known Little's law in Queueing theory [9].

5.4.4 Beyond the Mean Description

Fluctuations in the ASIP's occupancy vector grow as the system becomes larger. This Phenomenon is demonstrated in Figure 6 in which a homogenous ASIP system is simulated: For each site k, we have numerically calculated the steady state mean and standard deviation of the number of particles X_k present in the k^{th} site; the simulation vividly shows a *power law* growth of the standard deviation, as a function of the system's size (number of sites). This dramatic effect is facilitated by the batch service policy of the ASIP model which causes an "all or none" behavior which, in turn, leads to site occupancy fluctuations whose typical order of magnitude can be much larger than the mean occupancy itself. A mean based description fails to capture this complexity and other intricacies of the ASIP model. To fully capture the ASIP's statistics, we now turn to analyze its multidimensional probability distribution via probability generating functions.

5.5 PGF Dynamics

In this section we study the dynamics of the *Probability Generating Functions* (PGFs) of the random vectors $\mathbf{X}(t)$ $(t \geq 0)$ and $\mathbf{Y}(s)$ $(s = 1, 2, \cdots)$.

5.5.1 PGF Dynamics of $\mathbf{X}(t)$

The PGF of the random vector $\mathbf{X}(t)$ is given by

$$G_{\mathbf{X}}(t, z_1, z_2, \cdots, z_n) = \mathbf{E}\left[z_1^{X_1(t)} z_2^{X_2(t)} \cdots z_n^{X_n(t)}\right] \tag{36}$$

($|z_k| \leq 1$, $k = 1, \cdots, n$). Observe the system at times t and $t' = t + \Delta$ and use again the shorthand notation $\mathbf{X} = \mathbf{X}(t)$ and $\mathbf{X}' = \mathbf{X}(t')$. Conditioning on $\mathbf{X}$

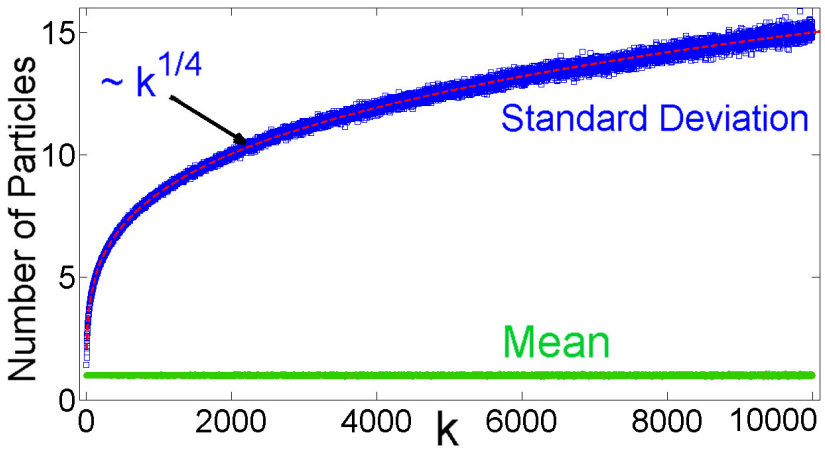

Figure 6: Large Fluctuations and the Emergence of Scaling Laws in the ASIP model. We have simulated a homogenous ASIP system with 10,000 sites, $\lambda = 1$ and $\mu_k = 1$ for $k = 1, ..., 10,000$. The mean and standard deviation in the number of particles at site k are plotted as a function of the site index. As expected, we find that regardless of the site index, on average each site is occupied by a single particle ($e_k = \frac{\lambda}{\mu_k} = 1$) . Conversely, the standard deviation in the number of particles exhibits a power law dependence on k and grows like $\sim k^{1/4}$ (dashed line is given by $N(k) = 1.5 \cdot k^{1/4}$). As fluctuation around the mean are typically much larger than the mean itself, it is clear that a mean based description is unable to capture the physics of large ASIP systems.

and utilizing the Markovian dynamics of Eq. (22) we have

$$\mathbf{E}\left[\prod_{k=1}^{n} z_k^{X_k'}\right] = \mathbf{E}\left[\mathbf{E}\left[\prod_{k=1}^{n} z_k^{X_k'} | \mathbf{X}\right]\right]$$

$$= \begin{cases} (1-(\lambda+\mu)\,\Delta)\,\mathbf{E}\left[\prod_{k=1}^{n} z_k^{X_k}\right] + (\lambda\Delta)\mathbf{E}\left[z_1 \prod_{k=1}^{n} z_k^{X_k}\right] \\ + \\ (\mu_1\Delta)\mathbf{E}\left[z_2^{X_1} \prod_{k\neq 1}^{n} z_k^{X_k}\right] + (\mu_2\Delta)\mathbf{E}\left[z_3^{X_2} \prod_{k\neq 2}^{n} z_k^{X_k}\right] \\ + \cdots + \\ (\mu_{n-1}\Delta)\mathbf{E}\left[z_n^{X_{n-1}} \prod_{k\neq (n-1)}^{n} z_k^{X_k}\right] + (\mu_n\Delta)\mathbf{E}\left[\prod_{k\neq n}^{n} z_k^{X_k}\right] \\ + \\ o(\Delta)\ . \end{cases} \tag{37}$$

Using the PGF notation of Eq. (36), Eq. (37) reads out as follows

$$
\begin{aligned}
&G_{\mathbf{X}}\left(t', z_1, z_2, \cdots, z_n\right) \\
&= \left\{ \begin{array}{l}
\left(1-(\lambda+\mu)\,\Delta\right) G_{\mathbf{X}}(t, z_1, z_2, z_{3,} \cdots, z_{n-1}, z_n) \\
+ \\
(\lambda\Delta) z_1 G_{\mathbf{X}}(t, z_1, z_2, z_{3,} \cdots, z_{n-1}, z_n) \\
+ \\
(\mu_1\Delta) G_{\mathbf{X}}(t, z_2, z_2, z_3, \cdots, z_{n-1}, z_n) \\
+ \\
(\mu_2\Delta) G_{\mathbf{X}}(t, z_1, z_3, z_3, \cdots, z_{n-1}, z_n) \\
+\cdots+ \\
(\mu_{n-1}\Delta) G_{\mathbf{X}}(t, z_1, z_2, z_{3,} \cdots, z_n, z_n) \\
+ \\
(\mu_n\Delta) G_{\mathbf{X}}(t, z_1, z_2, z_{3,} \cdots, z_{n-1}, 1) \\
+ \\
o(\Delta)\ .
\end{array} \right.
\end{aligned}
\tag{38}
$$

Rearranging the terms of Eq. (38), dividing by Δ, and taking $\Delta \to 0$ we conclude that

$$
\begin{aligned}
&\tfrac{\partial G_{\mathbf{X}}}{\partial t}\left(t, z_1, \cdots, z_n\right) \\
&= \left\{ \begin{array}{l}
\left[\lambda\left(z_1-1\right)-\mu\right] G_{\mathbf{X}}(t, z_1, z_2, z_{3,} \cdots, z_{n-1}, z_n) \\
+ \\
\mu_1 G_{\mathbf{X}}(t, z_2, z_2, z_3, \cdots, z_{n-1}, z_n) \\
+ \\
\mu_2 G_{\mathbf{X}}(t, z_1, z_3, z_3, \cdots, z_{n-1}, z_n) \\
+\cdots+ \\
\mu_{n-1} G_{\mathbf{X}}(t, z_1, z_2, z_{3,} \cdots, z_n, z_n) \\
+ \\
\mu_n G_{\mathbf{X}}(t, z_1, z_2, z_{3,} \cdots, z_{n-1}, 1)\ .
\end{array} \right.
\end{aligned}
\tag{39}
$$

Eq. (39) represents the 'PGF dynamics' of the random vector $\mathbf{X}\left(t\right)$. Namely, it transforms the Markovian dynamics of (22) to a differential equation of the form

$$
\frac{\partial G_{\mathbf{X}}}{\partial t}\left(t, \mathbf{z}\right) = \left[\mathcal{A} G_{\mathbf{X}}\right]\left(t, \mathbf{z}\right)\ , \tag{40}
$$

where $\mathbf{z} = \left(z_1, z_2, \cdots, z_n\right)$, and where $\mathcal{A}$ is an operator which acts only on the '$\mathbf{z}$-part' of the PGF $G_{\mathbf{X}}\left(t, \mathbf{z}\right)$.

5.5.2 PGF Dynamics of $\mathbf{Y}\left(s\right)$

The PGF of the random vector $\mathbf{Y}\left(s\right)$ is given by

$$
G_{\mathbf{Y}}\left(s, z_1, z_2, \cdots, z_n\right) = \mathbf{E}\left[z_1^{Y_1(s)} z_2^{Y_2(s)} \cdots z_n^{Y_n(s)}\right] \tag{41}
$$

($|z_k| \leq 1$, $k = 1, \cdots, n$). Observe the system at two consecutive s and $s' = s+1$ Poissonian events, and use again the shorthand notation $\mathbf{Y} = \mathbf{Y}(s)$ and $\mathbf{Y}' = \mathbf{Y}(s')$. Conditioning on $\mathbf{Y}$ and utilizing the 'law of motion' of Eq. (23) we have

$$\begin{aligned}
&\mathbf{E}\left[\prod_{k=1}^{n} z_k^{Y_k'}\right] = \mathbf{E}\left[\mathbf{E}\left[\prod_{k=1}^{n} z_k^{Y_k'} | \mathbf{Y}\right]\right] \\
&= \left\{\begin{array}{l}
\frac{\lambda}{\lambda+\mu}\mathbf{E}\left[z_1 \prod_{k=1}^{n} z_k^{Y_k}\right] + \frac{\mu_1}{\lambda+\mu}\mathbf{E}\left[z_2^{Y_1} \prod_{k\neq 1}^{n} z_k^{Y_k}\right] \\
+ \\
\frac{\mu_2}{\lambda+\mu}\mathbf{E}\left[z_3^{Y_2} \prod_{k\neq 2}^{n} z_k^{Y_k}\right] + \frac{\mu_3}{\lambda+\mu}\mathbf{E}\left[z_4^{Y_3} \prod_{k\neq 3}^{n} z_k^{Y_k}\right] \\
+ \cdots + \\
\frac{\mu_{n-1}}{\lambda+\mu}\mathbf{E}\left[z_n^{Y_{n-1}} \prod_{k\neq (n-1)}^{n} z_k^{Y_k}\right] + \frac{\mu_n}{\lambda+\mu}\mathbf{E}\left[\prod_{k\neq n}^{n} z_k^{Y_k}\right] \ .
\end{array}\right.
\end{aligned} \tag{42}$$

Using the PGF notation Eq. (41) and rearranging terms, Eq. (42) reads out as follows

$$\begin{aligned}
&G_{\mathbf{Y}}(s', z_1, \cdots, z_n) - G_{\mathbf{Y}}(s, z_1, \cdots, z_n) \\
&= \left\{\begin{array}{l}
\frac{\lambda(z_1-1)-\mu}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, \cdots, z_n) \\
+ \\
\frac{\mu_1}{\lambda+\mu} G_{\mathbf{Y}}(s, z_2, z_2, \cdots, z_n) \\
+ \\
\frac{\mu_2}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, z_3, z_3, \cdots, z_n) \\
+ \cdots + \\
\frac{\mu_{n-1}}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, \cdots, z_n, z_n) \\
+ \\
\frac{\mu_n}{\lambda+\mu} G_{\mathbf{Y}}(s, z_1, \cdots, z_{n-1}, 1) \ .
\end{array}\right.
\end{aligned} \tag{43}$$

Eq. (43) represents the 'PGF dynamics' of the random vector $\mathbf{Y}(s)$. Namely, it transforms the 'law of motion' of Eq. (43) to a difference equation of the form

$$G_{\mathbf{Y}}(s', \mathbf{z}) - G_{\mathbf{Y}}(s, \mathbf{z}) = [\mathcal{B} G_{\mathbf{Y}}](s, \mathbf{z}) \ , \tag{44}$$

where $\mathbf{z} = (z_1, z_2, \cdots, z_n)$, and where $\mathcal{B}$ is an operator which acts only the '$\mathbf{z}$-part' of the PGF $G_{\mathbf{Y}}(s, \mathbf{z})$.

5.5.3 Steady State

Consider now the ASIP model in *steady state*. The stochastic processes $(\mathbf{X}(t))_{t\geq 0}$ and $(\mathbf{Y}(s))_{s=1}^{\infty}$ are *stationary*, and hence their respective PGFs are time-homogeneous: $G_{\mathbf{X}}(t, \mathbf{z}) \equiv G_{\mathbf{X}}(\mathbf{z})$ $(t \geq 0)$ and $G_{\mathbf{Y}}(s, \mathbf{z}) \equiv G_{\mathbf{Y}}(\mathbf{z})$ $(s = 1, 2, \cdots)$. Substituting the time-homogeneous PGFs $G_{\mathbf{X}}(t, \mathbf{z}) \equiv G_{\mathbf{X}}(\mathbf{z})$ and $G_{\mathbf{Y}}(s, \mathbf{z}) \equiv G_{\mathbf{Y}}(\mathbf{z})$,

respectively, into Eqs. (39) and (43) yields the common equation

$$
\begin{aligned}
&\left[\lambda\left(1-z_1\right)+\mu\right] G(z_1, z_2, z_{3,} \cdots, z_{n-1}, z_n) \\
&= \begin{cases}
\mu_1 G(z_2, z_2, z_3, \cdots, z_{n-1}, z_n) \\
+ \\
\mu_2 G(z_1, z_3, z_3, \cdots, z_{n-1}, z_n) \\
+\cdots+ \\
\mu_{n-1} G(z_1, z_2, z_{3,} \cdots, z_n, z_n) \\
+ \\
\mu_n G(z_1, z_2, z_{3,} \cdots, z_{n-1}, 1) \ .
\end{cases}
\end{aligned} \tag{45}
$$

(where $G\left(\mathbf{z}\right)$ is the unknown function). Namely, both the PGFs $G_{\mathbf{X}}\left(\mathbf{z}\right)$ and $G_{\mathbf{Y}}\left(\mathbf{z}\right)$ are governed by Eq. (45).

Assuming that Eq. (45) admits a unique solution (we shall address the issue of uniqueness in Section 5.6) we obtain that: *In steady state, the distribution of the vector* $\mathbf{X}\left(t\right)$ *coincides with the distribution of the vector* $\mathbf{Y}\left(s\right)$. Namely, in steady state the ASIP displays the same statistics at arbitrary time epochs and at 'Poissonian events' time epochs. In the nomenclature of Queueing Theory such a phenomenon is termed PASTA: *Poisson Arrivals See Time Average* [9].

The PASTA phenomenon is a central concept in Queueing theory which implies that arriving customers find, on average, the same workload in the queueing system as an outside observer looking at the system at an arbitrary point in time. More precisely, the fraction of customers finding on arrival the system in some state S is exactly the same as the fraction of time the system is in state S. While well known results in queueing theory assert that the PASTA phenomenon holds for classes of systems with Poissonian arrivals (also known as $M/\cdot/\cdot$ queueing systems), this phenomenon does *not* hold for general systems. Indeed, even very simple queueing systems may fail to satisfy the PASTA phenomenon.

As an example consider the $D/D/1$ queueing system. In this system customers arrive to a service station with a single server in which they are processed according to their order of arrival. The customers' inter-arrival times and service times are deterministic. Let d_{arr} and d_{ser} denote, respectively, the deterministic inter-arrival and service times. Exactly every d_{arr} time units a new customer arrives at the service station, this customer must be served for exactly d_{ser} time units before leaving the system. Clearly the queue will explode if $d_{ser} > d_{arr}$, will be perfectly balanced if $d_{ser} = d_{arr}$, and will be stationary if $d_{ser} \leq d_{arr}$. If $d_{ser} < d_{arr}$ then the queue cycles will coincide with the customers' arrival epochs, the server will be busy for d_{ser} time units after arrival, and will be idle in the remaining $d_{arr} - d_{ser}$ time units. Clearly, arriving customers always observe an empty system (upon arrival). Hence the fraction of customers finding the system non-empty is zero. On the other hand the fraction of time the system is non-empty is d_{ser}/d_{arr}. The $D/D/1$ queueing model vividly exemplifies how the PASTA phenomenon can be violated even in very simple systems. On the other hand, the PASTA phenomenon can hold in complex processes such

as the running maxima of non-linear shot noise [62]. The fact that the PASTA phenomenon holds for all ASIP system is far from being trivial.

We now turn to describe the embedding phenomenon, another useful property of the ASIP. Consider two ASIP models: model (A) with m gates and parameters $\{\lambda, \mu_1, \cdots, \mu_m\}$, and model (B) with n gates and parameters $\{\lambda, \mu_1, \cdots, \mu_n\}$, where $m < n$. Eq. (45) implies the following embedding phenomenon: *The steady state distribution of model (A) coincides with the steady state distribution of the first m coordinates of model (B).* The derivation of the embedding phenomenon follows from substituting $z_{m+1} = \cdots = z_n = 1$ in Eq. (45). The intuitive understanding of the embedding phenomenon follows from the fact that in a ASIP model with n gates the operation of the first m gates ($k = 1, \cdots, m$) is indifferent to whatever happens in the following gates ($k = m+1, \cdots, n$). In other words, an observation of the first m sites in a ASIP model with n sites is indistinguishable from an observation of a ASIP model with m sites (and the same parameters).

5.6 Steady State Analysis

In this section we explore Eq. (45) governing the steady state PGF of the ASIP model.

5.6.1 Explicit Solution: $n = 1$

Consider the ASIP model with a single gate ($n = 1$). In this case Eq. (45) reduces to:

$$[\lambda(1 - z_1) + \mu_1] G(z_1) = \mu_1 G(1) \ . \tag{46}$$

Noting that $G(1) = 1$, and setting $p_1 = \mu_1 / (\mu_1 + \lambda)$, Eq. (46) implies that

$$G(z_1) = \frac{\mu_1}{\lambda(1 - z_1) + \mu_1} = \frac{p_1}{1 - (1 - p_1) z_1} \ . \tag{47}$$

The PGF of Eq. (47) characterizes the *Geometric Law* on the non-negative integers. Indeed, expanding both sides of Eq. (47) to power-series (in the variable z_1) yields the probability distribution

$$\Pr(X_1 = j) = \Pr(Y_1 = j) = (1 - p_1)^j p_1 \ , \tag{48}$$

$j = 0, 1, 2 \cdots$.

The probabilistic explanation of Eq. (48) is as follows: When $n = 1$, we can think about a 'competition' between two Poissonian processes — gate openings and particle arrivals . The Poissonian nature of these processes implies that the probability that the first Poissonian event is an arrival of a particle is $1 - p_1$. Similarly, the probability that the first Poissonian event is a gate opening is p_1. The memory-less property of the exponential distribution implies that in order for exactly k particles to leave the system at a gate opening moment, exactly k consecutive arrivals must be followed by a single gate opening. Hence the random variable Y_1 is geometrically distributed (on the non-negative integers)

with parameter p_1. As a result of the PASTA phenomenon described in the previous section, the distribution of the system vector at steady state is equal in law to the distribution of the system vector immediately after 'Poissonian events' - implying that X_1 coincides, in law, with Y_1.

5.6.2 Explicit Solution: $n = 2$

In this subsection we present, via the special case of $n = 2$, an iterative scheme for the solution of Eq. (45). In the basic step of the scheme, one uses Eq. (45) in order to obtain expressions for each of the generating functions that appear on its right-hand-side. Repeating the basic step, time and again, a 'branching tree structure' of generating functions forms. In this tree, each 'parent' generating function is expressed by a set of 'daughter' generating functions. As we shall demonstrate, the 'daughter' generating functions become somewhat simpler with every step of the scheme. Eventually, the 'daughter' generating functions become trivial — forming the 'leafs' of our branching tree. The scheme terminates once all 'daughter' generating functions turn into 'leafs'. The PGF is then obtained from 'transcending upwards' from the 'leafs' of the tree to its 'root'. At the 'root' an explicit, and by construction *unique*, expression for the PGF is attained.

Consider the ASIP model with two gates ($n = 2$). In this case Eq. (45) reduces to

$$\left[\lambda\left(1-z_1\right)+\mu_1+\mu_2\right]G(z_1,z_2)=\left\{\begin{array}{l}\mu_1 G(z_2,z_2)\\ +\\ \mu_2 G(z_1,1)\ .\end{array}\right. \tag{49}$$

Now, following the scheme's basic step, we iteratively apply Eq. (49) to the 'daughters' $G(z_2, z_2)$ and $G(z_1, 1)$.

For the 'daughter' $G(z_2, z_2)$ the basic step yields

$$\left[\lambda\left(1-z_2\right)+\mu_1+\mu_2\right]G(z_2,z_2)=\left\{\begin{array}{l}\mu_1 G(z_2,z_2)\\ +\\ \mu_2 G(z_2,1)\ .\end{array}\right. \tag{50}$$

from which we obtain that

$$G(z_2,z_2)=\frac{\mu_2}{\lambda\left(1-z_2\right)+\mu_2}G(z_2,1)\ . \tag{51}$$

In turn, for the 'daughter' $G(z_2, 1)$ Eq. (49) yields

$$\left[\lambda\left(1-z_2\right)+\mu_1+\mu_2\right]G(z_2,1)=\left\{\begin{array}{l}\mu_1 G(1,1)\\ +\\ \mu_2 G(z_2,1)\ .\end{array}\right. \tag{52}$$

from which we obtain that

$$G(z_2,1)=\frac{\mu_1}{\lambda\left(1-z_2\right)+\mu_1}G(1,1)\ . \tag{53}$$

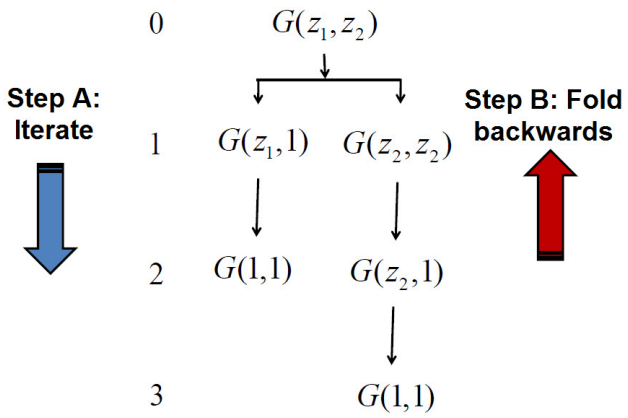

Figure 7: Schematic illustration of the iterative solution of Eq. (45) for $n = 2$. Step A: Eq. (45) is iterated repeatedly, in a branching-tree structure, till reaching the 'leafs' G(1,1)=1. Step B: The tree is 'folded back' yielding the value of the 'root' $G(z_1, z_2)$ (Eq. (57)).

For the 'daughter' $G(z_1, 1)$ the iteration yields

$$[\lambda(1 - z_1) + \mu_1 + \mu_2] G(z_1, 1) = \begin{cases} \mu_1 G(1,1) \\ + \\ \mu_2 G(z_1, 1) \ . \end{cases} \tag{54}$$

from which we obtain that

$$G(z_1, 1) = \frac{\mu_1}{\lambda(1 - z_1) + \mu_1} G(1, 1) \ . \tag{55}$$

The 'leafs' of our 'tree' are characterized by the PGF $G(1, 1)$ which trivially equals unity. Hence, setting $G(1, 1) = 1$ in the 'leafs' — Eqs. (53) and (55) — yields the 'daughters' $G(z_2, 1)$ and $G(z_1, 1)$. Substituting the 'daughter' $G(z_2, 1)$ into Eq. (51) yields the 'daughter'

$$G(z_2, z_2) = \frac{\mu_2}{\lambda(1 - z_2) + \mu_2} \frac{\mu_1}{\lambda(1 - z_2) + \mu_1} \ . \tag{56}$$

Finally, substituting the 'daughters' $G(z_2, z_2)$ and $G(z_1, 1)$ into Eq. (49) yields the 'root'

$$G(z_1, z_2) = \begin{cases} \frac{\mu_1^2 \mu_2}{[\lambda(1-z_2)+\mu_2][\lambda(1-z_2)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]} \\ + \\ \frac{\mu_1 \mu_2}{[\lambda(1-z_1)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]} \ . \end{cases} \tag{57}$$

Summarizing, we have found that for $n = 2$ the scheme terminates after two iterations. The result is a tree-like structure whose 'leafs' are trivial constants all equal to unity. Knowing the constants that stand in the base of the tree,

we are able to calculate the functions that occupy the second lowest level. The PGF $G(z_1, z_2)$ was computed by iterating this procedure, i.e., by using known functions at the 'current knowledge level' of the tree in order to compute the functions at the next level. A 'tree sketch' of the solution steps for the ASIP model with two gates ($n = 2$) is depicted in Figure 7.

5.6.3 Explicit Solution: $n = 3$

The iterative scheme described in the previous subsection applies, in theory, to the ASIP model with an arbitrary number of gates. In practice however, the solution's complexity increases rapidly with the number of gates n. Thus, effectively, for large n, the PGF of the ASIP model is not tractable. To illustrate just how dramatically the solution complexity increases — consider the ASIP model with three gates ($n = 3$). In the Appendix to this chapter, we show that the expression for $G(z_1, z_2, z_3)$ is given by

$$
\begin{aligned}
&G(z_1, z_2, z_3) \\
&= \left\{ \begin{array}{l}
\frac{\mu_1^3\mu_2^2\mu_3/[\lambda(1-z_2)+\mu_1+\mu_3]}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_3)+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1]} \\
+ \\
\frac{\mu_1^3\mu_2^2\mu_3/[\lambda(1-z_2)+\mu_1+\mu_3]}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1][\lambda(1-z_2)+\mu_1+\mu_2]} \\
+ \\
\frac{\mu_1^2\mu_2^2\mu_3/[\lambda(1-z_2)+\mu_1+\mu_3]}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_2)+\mu_1][\lambda(1-z_2)+\mu_1+\mu_2]} \\
+ \\
\frac{\mu_1^2\mu_2\mu_3}{[\lambda(1-z_2)+\mu_2+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_2)+\mu_2][\lambda(1-z_2)+\mu_1]} \\
+ \\
\frac{\mu_1^2\mu_2^2\mu_3}{[\lambda(1-z_1)+\mu_1+\mu_3][\lambda(1-z_3)+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1]} \\
+ \\
\frac{\mu_1^2\mu_2^2\mu_3/[\lambda(1-z_1)+\mu_1+\mu_3]}{[\lambda(1-z_1)+\mu][\lambda(1-z_3)+\mu_2][\lambda(1-z_3)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]} \\
+ \\
\frac{\mu_1\mu_2^2\mu_3}{[\lambda(1-z_1)+\mu_1+\mu_3][\lambda(1-z_1)+\mu][\lambda(1-z_1)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]} \\
+ \\
\frac{\mu_1^2\mu_2\mu_3}{[\lambda(1-z_1)+\mu][\lambda(1-z_2)+\mu_2][\lambda(1-z_2)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]} \\
+ \\
\frac{\mu_1\mu_2\mu_3}{[\lambda(1-z_1)+\mu][\lambda(1-z_1)+\mu_1][\lambda(1-z_1)+\mu_1+\mu_2]} \ .
\end{array} \right.
\end{aligned}
\tag{58}
$$

Eq. (58) well exemplifies the intrinsic complexity of the ASIP model. A 'tree sketch' of the solution steps for the ASIP model with three gates ($n = 3$) is depicted in Figure 8.

We note that at first glance it might seem possible to derive the steady-state *marginal* distributions of the random variables $\{X_1(t), \cdots, X_n(t)\}$ iteratively. Namely, to establish a recursion equation relating the PGF of $X_k(t)$ to the PGF of $X_{k-1}(t)$, and then solve it. However, the random variables

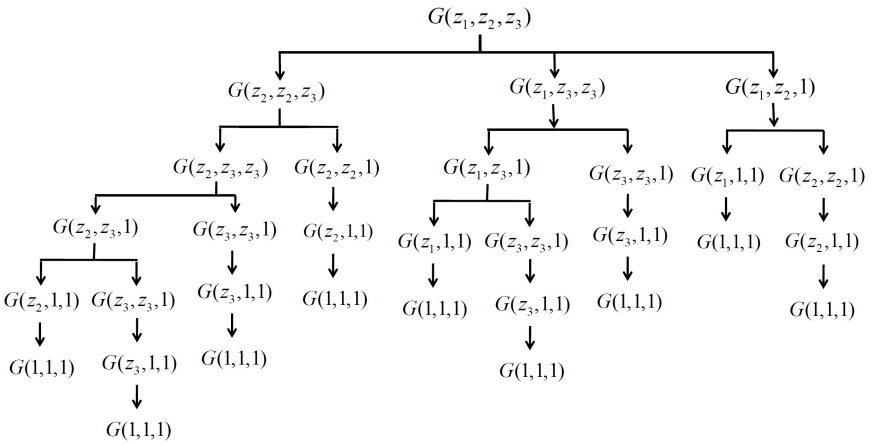

Figure 8: Schematic illustration of the iterative solution of Eq. (45) for $n = 3$. Step A: Eq. (45) is iterated repeatedly, in a branching-tree structure, till reaching the 'leafs' G(1,1,1)=1. Step B: The tree is 'folded back' yielding the value of the 'root' $G(z_1, z_2, z_3)$ (Eq. (58)).

$\{X_1(t), \cdots, X_k(t)\}$ are correlated, and when trying to establish the aforementioned recursion equation several complications arise. An exact solution for the marginals can be obtained nevertheless and we return to this topic in Chapter 8.

5.7 Load Analysis

In Sections 5.5 and 5.6 we studied the PGF dynamics and steady state solution of the random vectors $\mathbf{X}(t)$ and $\mathbf{Y}(s)$, and demonstrated intractability for large n. In this section we study the *load* of the ASIP model: the total number of particles in the system — given by the sum-of-coordinates of the random vectors $\mathbf{X}(t)$ and $\mathbf{Y}(s)$. Surprisingly so, the load turns out to be analytically tractable.

5.7.1 Load Analysis of $\mathbf{X}(t)$

Let $X_{(k)}(t)$ denote the total number of particles present, at an arbitrary time t ($t \geq 0$), in the first k sites. The random variable $X_{(k)}(t)$ is the sum of the first k coordinates of the random vector $\mathbf{X}(t)$, i.e., $X_{(k)}(t) = X_1(t) + \cdots + X_k(t)$ ($k = 1, \cdots, n$).

Observe the system at times t and $t' = t + \Delta$, and use the shorthand notation $X_{(k)} = X_{(k)}(t)$ and $X'_{(k)} = X_{(k)}(t')$ ($k = 1, \cdots, n$). Eq. (22) implies that the stochastic connection between the random sums $X_{(k)}$ and $X'_{(k)}$ — characterizing

the Markovian dynamics of the stochastic process $\left(X_{(k)}(t)\right)_{t\geq 0}$ — is given by:

$$X'_{(k)} = \begin{cases} X_{(k)} & \text{w.p.} \quad 1-(\lambda+\mu_k)\,\Delta+o(\Delta)\ , \\ X_{(k)}+1 & \text{w.p.} \quad \lambda\Delta+o(\Delta)\ , \\ X_{(k-1)} & \text{w.p.} \quad \mu_k\Delta+o(\Delta)\ . \end{cases} \tag{59}$$

Eq. (59) follows from considering the events that may take place — and result in a change $X_{(k)} \mapsto X'_{(k)} \neq X_{(k)}$ — within the time interval $(t,t']$. There are exactly two such events, and we label them according to the Poisson processes inducing them: (0) the arrival of a particle to the first site — occurring with probability $\lambda\Delta+o(\Delta)$ — in which case $X_{(k)} \mapsto X'_{(k)} = X_{(k)}+1$; (k) opening of the k^{th} gate — occurring with probability $\mu_k\Delta+o(\Delta)$ — in which case $X_{(k)} \mapsto X'_{(k)} = X_{(k-1)}$. The first line on the right-hand-side of Eq. (59) represents the scenario in which no event takes place — which occurs with the complementary probability $1-(\lambda+\mu_k)\,\Delta+o(\Delta)$.

Let

$$G_{X_{(k)}}(t,z) = \mathbf{E}\left[z^{X_{(k)}(t)}\right] \tag{60}$$

($|z| \leq 1$) denote the PGF of the random sum $X_{(k)}(t)$. Setting $z_1 = \cdots = z_k = z$ and $z_{k+1} = \cdots = z_n = 1$, and noting that by definition $G_{X_{(k)}}(t,z) = G_{\mathbf{X}}(t,z,\cdots,z,1,\cdots,1)$, Eq. (39) yields

$$\frac{\partial G_{X_{(k)}}}{\partial t}(t,z) = \begin{cases} \left[\lambda(z-1)-\mu\right] G_{X_{(k)}}(t,z) \\ + \\ \mu_1 G_{X_{(k)}}(t,z) \\ +\cdots+ \\ \mu_{k-1} G_{X_{(k)}}(t,z) \\ + \\ \mu_k G_{X_{(k-1)}}(t,z) \\ + \\ \mu_{k+1} G_{X_{(k)}}(t,z) \\ +\cdots+ \\ \mu_{n-1} G_{X_{(k)}}(t,z) \\ + \\ \mu_n G_{X_{(k)}}(t,z)\ . \end{cases} \tag{61}$$

Eq. (61), in turn, implies that the 'PGF dynamics' of the random sum $X_{(k)}(t)$ are given by

$$\begin{aligned} &\tfrac{\partial G_{X_{(k)}}}{\partial t}(t,z) \\ &= \left[\lambda(z-1)-\mu_k\right] G_{X_{(k)}}(t,z) + \mu_k G_{X_{(k-1)}}(t,z)\ . \end{aligned} \tag{62}$$

5.7.2 Load Analysis of $\mathbf{Y}(s)$

Let $Y_{(k)}(s)$ denote the total number of particles present, immediately after the s^{th} Poissonian event ($s = 1, 2, ...$), in the first k sites. The random variable $Y_{(k)}(s)$ is the sum of the first k coordinates of the random vector $\mathbf{Y}(s)$, i.e., $Y_{(k)}(s) = Y_1(s) + \cdots + Y_k(s)$ $(k = 1, \cdots, n)$.

Observe the system at two consecutive s and $s' = s+1$ Poissonian events, and use the shorthand notation $Y_{(k)} = Y_{(k)}(s)$ and $Y'_{(k)} = Y_{(k)}(s')$ $(k = 1, \cdots, n)$. Eq. (23) implies that the stochastic connection between the random sums $Y_{(k)}$ and $Y'_{(k)}$ — characterizing the 'law of motion' of the stochastic process $\left(Y_{(k)}(s)\right)_{s=1}^{\infty}$ — is given by:

$$Y'_{(k)} = \begin{cases} Y_{(k)} & \text{w.p.} \quad \frac{\mu - \mu_k}{\lambda + \mu} \ , \\ Y_{(k)} + 1 & \text{w.p.} \quad \frac{\lambda}{\lambda + \mu} \ , \\ Y_{(k-1)} & \text{w.p.} \quad \frac{\mu_k}{\lambda + \mu} \ . \end{cases} \tag{63}$$

Eq. (63) follows from considering the events that result in a change $Y_{(k)} \mapsto Y'_{(k)} \neq Y_{(k)}$. There are exactly two such events, and we label them according to the Poisson processes inducing them: (0) the arrival of a particle to the first site — occurring with probability $\lambda/(\lambda + \mu)$ — in which case $Y_{(k)} \mapsto Y'_{(k)} = Y_{(k)} + 1$; (k) opening of the k^{th} gate — occurring with probability $\mu_k/(\lambda + \mu)$ — in which case $Y_{(k)} \mapsto Y'_{(k)} = Y_{(k-1)}$. The first line on the right-hand-side of Eq. (63) represents the scenario in which a gate other than the k^{th} gate opens — which occurs with the complementary probability $(\mu - \mu_k)/(\lambda + \mu)$.

Let

$$G_{Y_{(k)}}(s, z) = \mathbf{E}\left[z^{Y_{(k)}(s)}\right] \tag{64}$$

$(|z| \leq 1)$ denote the PGF of the random sum $Y_{(k)}(s)$. Setting $z_1 = \cdots = z_k = z$ and $z_{k+1} = \cdots = z_n = 1$, and noting that by definition $G_{Y_{(k)}}(s, z) = G_{\mathbf{Y}}(s, z, \cdots, z, 1, \cdots, 1)$, Eq. (43) yields

$$G_{Y_{(k)}}(s', z) - G_{Y_{(k)}}(s, z) = \begin{cases} \frac{\lambda(z-1) - \mu}{\lambda + \mu} G_{Y_{(k)}}(s, z) \\ + \\ \frac{\mu_1}{\lambda + \mu} G_{Y_{(k)}}(s, z) \\ + \cdots + \\ \frac{\mu_{k-1}}{\lambda + \mu} G_{Y_{(k)}}(s, z) \\ + \\ \frac{\mu_k}{\lambda + \mu} G_{Y_{(k-1)}}(s, z) \\ + \\ \frac{\mu_{k+1}}{\lambda + \mu} G_{Y_{(k)}}(s, z) \\ + \cdots + \\ \frac{\mu_{n-1}}{\lambda + \mu} G_{Y_{(k)}}(s, z) \\ + \\ \frac{\mu_n}{\lambda + \mu} G_{Y_{(k)}}(s, z) \ . \end{cases} \tag{65}$$

Eq. (65), in turn, implies that the 'PGF dynamics' of the random sum $Y_{(k)}(s)$ is given by

$$\begin{aligned} & G_{Y_{(k)}}(s',z) - G_{Y_{(k)}}(s,z) \\ & = \tfrac{\lambda(z-1)-\mu_k}{\lambda+\mu} G_{Y_{(k)}}(s,z) + \tfrac{\mu_k}{\lambda+\mu} G_{Y_{(k-1)}}(s,z) \ . \end{aligned} \tag{66}$$

5.7.3 Steady State

Consider now the ASIP model in *steady state*. In steady state the stochastic processes $\left(X_{(k)}(t)\right)_{t\geq 0}$ and $\left(Y_{(k)}(s)\right)_{s=1}^{\infty}$ are *stationary*, and hence their respective PGFs are time-homogeneous: $G_{X_{(k)}}(t,z) \equiv G_{X_{(k)}}(z)$ $(t \geq 0)$ and $G_{Y_{(k)}}(s,z) \equiv G_{Y_{(k)}}(z)$ $(s = 1,2,\cdots)$. Substituting the time-homogeneous PGFs $G_{X_{(k)}}(t,z) \equiv G_{X_{(k)}}(z)$ and $G_{Y_{(k)}}(s,z) \equiv G_{Y_{(k)}}(z)$, respectively, into Eqs. (62) and (66) yields the common equation

$$G_k(z) = \frac{\mu_k}{\mu_k + \lambda(1-z)} G_{k-1}(z) \tag{67}$$

$(k = 1,\cdots,n)$. Namely, both the PGFs $G_{X_{(k)}}(z)$ and $G_{Y_{(k)}}(z)$ are governed by Eq. (67).

Note that $X_{(1)}(t) = X_1(t)$ and $Y_{(1)}(s) = Y_1(s)$ and hence the PGF $G_1(z)$ is given by Eq. (47). Using the 'initial condition' $G_1(z)$ and iterating Eq. (67) we obtain that

$$\begin{aligned} & \mathbf{E}\left[z^{X_{(k)}(t)}\right] = \mathbf{E}\left[z^{Y_{(k)}(s)}\right] \\ & = \tfrac{\mu_1}{\mu_1+\lambda(1-z)} \cdots \tfrac{\mu_k}{\mu_k+\lambda(1-z)} \\ & = \tfrac{p_1}{1-(1-p_1)z} \cdots \tfrac{p_k}{1-(1-p_k)z} \ , \end{aligned} \tag{68}$$

where $p_k = \mu_k/(\mu_k+\lambda)$ $(k = 1,\cdots,n)$. As the 'initial condition' implies — in the case $k = 1$, Eq. (68) coincides with Eq. (47). Interestingly, for $k > 1$, Eq. (68) attains a *product-form representation*. This product-form implies that both $X_{(k)}(t)$ and $Y_{(k)}(t)$ are characterized by the following *stochastic decomposition*: The random variables $X_{(k)}(t)$ and $Y_{(k)}(t)$ are equal, in law, to the total number of particles in k *independent* and *single gated* ASIP systems with respective parameters $(\lambda,\mu_1),\cdots,(\lambda,\mu_k)$.

The PGF $G(z) = p/[1-(1-p)z]$ characterizes a *Geometric Law* on the non-negative integers — which, in turn, has mean $\frac{1-p}{p}$ and variance $\frac{1-p}{p^2}$. Combining this fact together with the aforementioned stochastic representation, we obtain that the mean and variance of the random variables $X_{(k)}(t)$ and $Y_{(k)}(t)$ are given, respectively, by

$$\begin{aligned} & \mathbf{E}\left[X_{(k)}(t)\right] = \mathbf{E}\left[Y_{(k)}(s)\right] \\ & = \lambda\left(\tfrac{1}{\mu_1} + \cdots + \tfrac{1}{\mu_k}\right) \end{aligned} \tag{69}$$

and

$$\begin{aligned}\mathbf{Var}\left[X_{(k)}(t)\right] &= \mathbf{Var}\left[Y_{(k)}(s)\right] \\ &= \lambda\left(\tfrac{\mu_1+\lambda}{\mu_1^2}+\cdots+\tfrac{\mu_k+\lambda}{\mu_k^2}\right) .\end{aligned} \tag{70}$$

We emphasize that Eq. (68) implies that the distribution of $X_{(n)}(t)$ and $Y_{(n)}(s)$ is *independent of the order of the gates*. Consequently, permuting the gates (each gate carrying its own opening rate with it) has no effect on the distribution of the ASIP load. Thus, from a load-perspective, the ASIP model is invariant with respect to gate permutations.

5.8 Load Optimization

To design an efficient ASIP system one would like to minimize the system's load, i.e., to minimize the number of particles 'in process' termed "Work-In-Process" (WIP in production models [63]). In this section we explore the optimization of the ASIP's load. In what follows we consider as given the 'exogenous' inflow rate λ, and optimize the 'endogenous' service rates $\{\mu_1, \mu_2, \cdots, \mu_n\}$.

5.8.1 Optimality

We begin with the *combinatorial optimization* of the ASIP model. Namely, given a collection of n gates — each with its own service rate — we seek an ordering of gates that renders a target functional optimal. As explained in the previous section, the distribution of the ASIP's load is invariant with respect to gate permutations. Hence, for any target functional based on the ASIP's load-distribution optimization is trivial: all gate permutations yield the same target-functional value.

Let us turn now to examine *constrained optimization* of the ASIP model. To that end we consider four optimization problems in which we seek to minimize a target functional based on the ASIP's load-distribution, subject to a given constraint:

1. **Minimization of the load-mean subject to a given cumulative service rate**. Assume that the cumulative service rate μ is fixed and constant. Here we seek an optimal allocation of the cumulative service rate μ to the different gates — the goal being a minimal load-mean. Recalling Eq. (35) which asserts that the load-mean is given by the product $\lambda\mathbf{E}[T]$, we note that the minimization of the load-mean is equivalent to the minimizing of the traversal time. Applying Eq. (69) we obtain the constrained optimization problem

$$\begin{cases} \quad min\left\{\lambda\left(\frac{1}{\mu_1}+\cdots+\frac{1}{\mu_n}\right)\right\} \\ \text{s.t.} \\ \quad \mu_1+\cdots+\mu_n=\mu\ . \end{cases} \tag{71}$$

2. **Minimization of the load-variance subject to a given cumulative service rate**. Assume that the cumulative service rate μ is fixed and constant. Here we seek an optimal allocation of the cumulative service rate μ to the different gates — the goal being a minimal load-variance. Applying Eq. (70) we obtain the constrained optimization problem

$$\begin{cases} \quad min \left\{ \lambda \left(\frac{\mu_1+\lambda}{\mu_1^2} + \cdots + \frac{\mu_n+\lambda}{\mu_n^2} \right) \right\} \\ \text{s.t.} \\ \quad \mu_1 + \cdots + \mu_n = \mu \ . \end{cases} \tag{72}$$

3. **Minimization of the load-variance subject to a given load-mean**. Assume the load-mean is predetermined to equal the value v (alternatively, assume that the traversal time is predetermined to equal the value v/λ). Here we seek the optimal service rates that render the load-variance minimal. Applying Eqs. (69) and (70) we obtain the constrained optimization problem

$$\begin{cases} \quad min \left\{ \lambda \left(\frac{\mu_1+\lambda}{\mu_1^2} + \cdots + \frac{\mu_n+\lambda}{\mu_n^2} \right) \right\} \\ \text{s.t.} \\ \quad \lambda \left(\frac{1}{\mu_1} + \cdots + \frac{1}{\mu_n} \right) = v \ . \end{cases} \tag{73}$$

   This optimization problem is analogous to the 'Markowitz optimization' of financial portfolios — in which one seeks to minimize the portfolio variance, subject to a predetermined portfolio mean [64].

4. **Maximization of the zero-load probability subject to a given cumulative service rate**. Assume that the cumulative service rate is fixed and constant. Here we seek an optimal allocation of the cumulative service rate to the different gates — the goal being a maximal zero-load probability $\Pr(X_{(n)}(t) = 0)$. This zero-load probability is attained by setting $z = 0$ into the PGF of the load $X_{(n)}(t)$. Setting $z = 0$ into the right-hand-side of Eq. (68) we obtain the constrained optimization problem

$$\begin{cases} \quad max \left\{ \frac{\mu_1}{\mu_1+\lambda} \cdots \frac{\mu_n}{\mu_n+\lambda} \right\} \\ \text{s.t.} \\ \quad \mu_1 + \cdots + \mu_n = \mu \ . \end{cases} \tag{74}$$

   Note that the constrained optimization problem Eq. (74) is equivalent to the constrained optimization problem

$$\begin{cases} \quad min \left\{ ln \left(1 + \frac{\lambda}{\mu_1} \right) + \cdots + ln \left(1 + \frac{\lambda}{\mu_n} \right) \right\} \\ \text{s.t.} \\ \quad \mu_1 + \cdots + \mu_n = \mu \ . \end{cases} \tag{75}$$

The four aforementioned optimization problems admit the general form

$$\begin{cases} \quad min \left\{ f(x_1) + \cdots + f(x_n) \right\} \\ \text{s.t.} \\ \quad x_1 + \cdots + x_n = c \ , \end{cases} \tag{76}$$

where $f(x)$ is a convex function and the variables are positive valued: $x_1, \cdots, x_n > 0$. Indeed: (1) in the first problem $x_k = \mu_k$, $c = \mu$, and $f(x) = \lambda/x$; (2) in the second problem $x_k = \mu_k$, $c = \mu$, and $f(x) = (\lambda/x) + (\lambda/x)^2$; (3) in the third problem $x_k = 1/\mu_k$, $c = v/\lambda$, and $f(x) = (\lambda x) + (\lambda x)^2$; (4) in the fourth problem $x_k = \mu_k$, $c = \mu$, and $f(x) = \ln(1 + \lambda/x)$. The Lagrange function corresponding to the optimization problem of Eq. (76) is given by

$$L(x_1, \cdots, x_n; \theta) = \left(\sum_{k=1}^{n} f(x_k) \right) + \theta \left(c - \sum_{k=1}^{n} x_k \right) . \tag{77}$$

Differentiating the Lagrange function with respect to the variable x_k and equating the partial derivative to zero yields the equation

$$f'(x_k) = \theta . \tag{78}$$

Now, since the function $f(x)$ is convex ($f''(x) > 0$) its derivative $f'(x)$ is monotone increasing. This implies that Eq. (78) admits a unique solution — which, in turn, implies that the unique critical point of the Lagrange function satisfies $x_1 = \cdots = x_n$. Since the target function $\sum_{k=1}^{n} f(x_k)$ is convex, and the constraint $\sum_{k=1}^{n} x_k = c$ is linear, we conclude that [65]: the global minimum of the optimization problem (76) is given by $x_1 = \cdots = x_n = c/n$.

Thus, the solution to all four aforementioned optimization problems turns out to be an *homogenous ASIP system* — with service rates $\mu_1 = \mu_2 = \cdots = \mu_n$. This optimization conclusion highlights the importance of homogenous ASIP systems within the class of general ASIP systems. The optimality of the homogenous solution is illustrated graphically in Figure 9. Panels A-C are associated with the optimization problems presented in Eqs. (71), (72) and (74) respectively. In all three panels, results are shown for ASIP systems with 25 gates ($n = 25$) and an inflow rate of $\lambda = 1$. The constraint parameters, μ in Eqs. (71) and (72) and v in Eq. (74) are taken to equal 25 ($\mu = v = 25$). The optimal solution under these conditions is identical for all three problems and is given by $\mu_1 = \mu_2 = \cdots = \mu_{25} = 1$. The value of the target function, evaluated at randomly drawn rate vectors $(\mu_1, \cdots, \mu_{25})$, is plotted vs. the Euclidean distance of these vectors from the optimal rate vector $(1, 1, \cdots, 1)$. The optimality of the latter is clearly visible. In each panel, rate vectors are randomly drawn 25,000 times in the two following methods: 'Gaussian Sampling' and 'Uniform Sampling'. In the Gaussian Sampling, Gaussian noise is added to the optimal rate vector. This vector is then normalized to form a rate vector that complies with the problem constraints. In the Uniform Sampling, the interval $[0, 25]$ is dissected into 25 segments by randomly drawing 24 numbers from a uniform distribution over that interval. The lengths of these segments are then taken to represent the rate vector (in the case of panels A and B) or the inverse rate vector $(\frac{1}{\mu_1}, \cdots, \frac{1}{\mu_{25}})$ in the case of panel C.

5.8.2 Deviations from Optimality and Bottlenecks

Having concluded that homogenous ASIP systems are optimal we turn to discuss deviations from optimality. Of particular interest is the sensitivity of the target

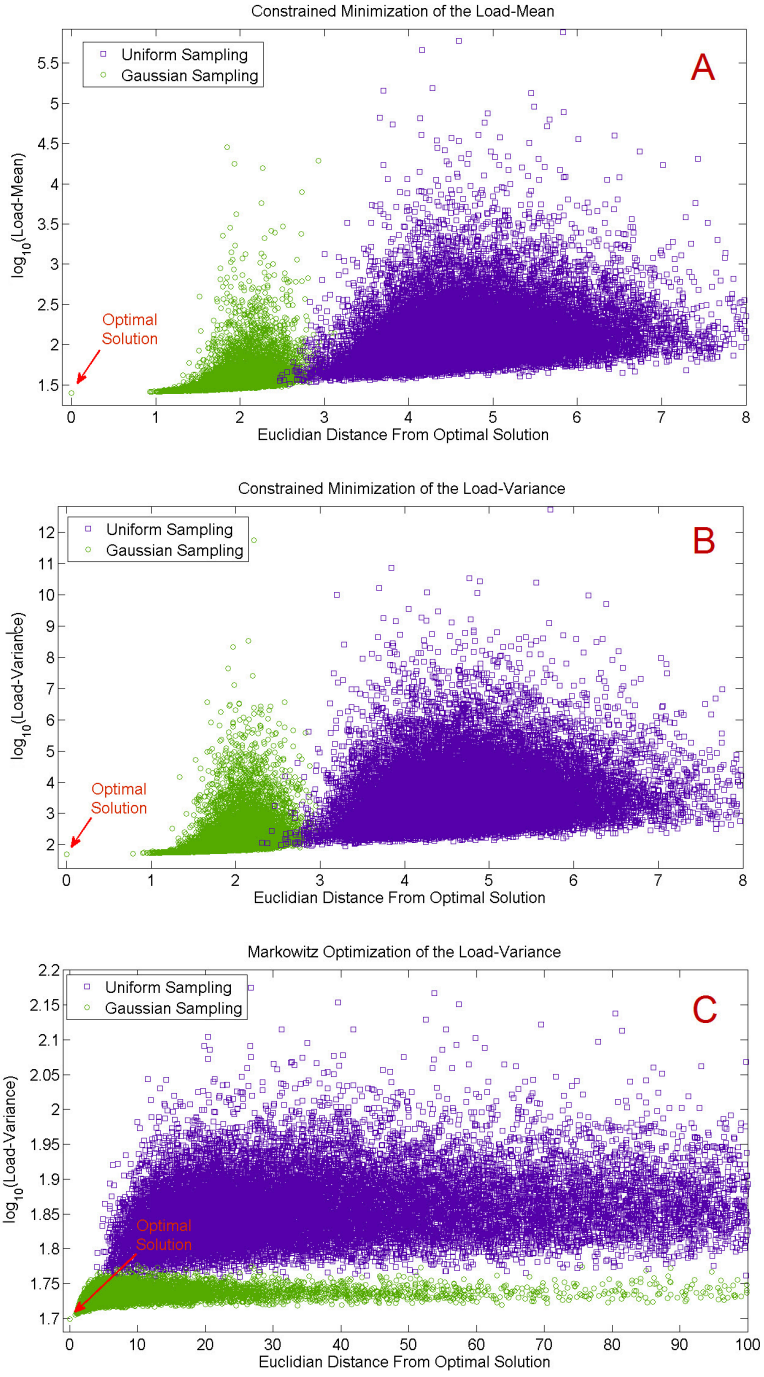

Figure 9: Optimality of homogeneous ASIP systems.

function to small changes in the optimal service rates vector. To this end we find it useful to borrow the "bottleneck" concept from the ASEP nomenclature [66]. Bottlenecks are sites where the hopping rate of particles is reduced compared to the rest of the system. In the ASEP the main effect of bottlenecks is to decrease the current (or flow) throw the system [67]. In the ASIP the steady state flow of particles is always given by λ and is hence independent of the service rates $\{\mu_1, \mu_2, \cdots, \mu_n\}$. Interestingly, bottlenecks are nevertheless useful in understanding deviations from optimality since both the load-mean and load-variance of an ASIP system are sensitive to their existence.

When it comes to the sensitivity of the target function to perturbations around the optimal solution, the third panel of Figure 9 shows a strikingly different behavior compared to the first two. While in the first two panels small deviations from the optimal rates vector may change the target function by orders of magnitude, this is hardly the case in the third panel.

This phenomenon can easily be understood by noting that in the first two optimization problems discussed above, the given cumulative service rate constraint does not impose a lower bound on the service rates in the system. This constraint can therefore be satisfied even in the presence of a site whose service rate is infinitesimally small. A single bottleneck (or defect) within an otherwise (almost) homogenous ASIP system will result in nothing but a slight deviation from the optimal solution. However, since the load-mean and load-variance are highly sensitive to the existence of bottlenecks, the impact on the target function will be tremendous.

Situation is considerably different when the given cumulative service rate constraint is replaced by a given load-mean constraint as is done in the third optimization problem above. The latter imposes a lower bound on the service rates in the system. Moreover, in order to satisfy the constraint, the existence of a bottleneck forces the allocation of extremely high service rates to many other sites. And so, the impact of bottlenecks on the target function is both limited and, when substantial, accompanied by a discernible deviation from the optimal solution.

5.9 Conclusions

In this chapter we analyzed the basic features of the ASIP and obtained the following results: (i) explicit evolution equations for the mean and PGF; (ii) explicit solution of the mean in steady-state; (iii) explicit equations for the PGF in steady-state; (iv) explicit solution of the steady-state PGF for small systems ($n = 1, 2, 3$), and an iterative scheme for the computation of the steady-state PGF for systems of arbitrary size; (v) explicit solution of the mean, variance, and PGF of the load in steady-state; (vi) explicit solutions of various load-optimization problems — rendering homogenous ASIP models optimal. In the following chapter, we will examine the behavior of the ASIP under various limiting regimes.

5.10 Appendix

Here we solve Eq. (45) for the ASIP with three gates $n = 3$. In this case Eq. (45) reduces to

$$[\lambda(1-z_1)+\mu]\,G(z_1,z_2,z_3) = \begin{cases} \mu_1 G(z_2,z_2,z_3) \\ + \\ \mu_2 G(z_1,z_3,z_3) \\ + \\ \mu_3 G(z_1,z_2,1)\ . \end{cases} \tag{79}$$

Now, following the scheme's basic step, we iteratively apply Eq. (79) to the 'daughters' $G(z_2,z_2,z_3)$, $G(z_1,z_3,z_3)$, and $G(z_1,z_2,1)$.

From the embedding property (see end of Section 5.5), the 'daughter' $G(z_1,z_2,1)$ is equal to $G(z_1,z_2)$ and is hence known and given by Eq. (57). For the 'daughter' $G(z_1,z_3,z_3)$ the basic step yields

$$[\lambda(1-z_1)+\mu]\,G(z_1,z_3,z_3) = \begin{cases} \mu_1 G(z_3,z_3,z_3) \\ + \\ \mu_2 G(z_1,z_3,z_3) \\ + \\ \mu_3 G(z_1,z_3,1) \end{cases} \tag{80}$$

from which we obtain

$$[\lambda(1-z_1)+\mu_1+\mu_3]\,G(z_1,z_3,z_3) = \begin{cases} \mu_1 G(z_3,z_3,z_3) \\ + \\ \mu_3 G(z_1,z_3,1)\ . \end{cases} \tag{81}$$

Again, the 'daughter' $G(z_1,z_3,1)$ is known and given by Eq. (57). For the 'daughter' $G(z_3,z_3,z_3)$ the basic step yields

$$[\lambda(1-z_3)+\mu]\,G(z_3,z_3,z_3) = \begin{cases} \mu_1 G(z_3,z_3,z_3) \\ + \\ \mu_2 G(z_3,z_3,z_3) \\ + \\ \mu_3 G(z_3,z_3,1) \end{cases} \tag{82}$$

from which we obtain

$$[\lambda(1-z_3)+\mu_3]\,G(z_3,z_3,z_3) = \mu_3 G(z_3,z_3,1)\ . \tag{83}$$

We conclude that

$$G(z_1,z_3,z_3) = \begin{cases} \frac{\mu_1\mu_3 G(z_3,z_3,1)}{(\lambda(1-z_1)+\mu_1+\mu_3)(\lambda(1-z_3)+\mu_3)} \\ + \\ \frac{\mu_3 G(z_1,z_3,1)}{\lambda(1-z_1)+\mu_1+\mu_3}\ . \end{cases} \tag{84}$$

We now return to the 'daughter' $G(z_2, z_2, z_3)$, applying the basic step yields

$$[\lambda(1-z_2)+\mu]\,G(z_2,z_2,z_3) = \begin{cases} \mu_1 G(z_2,z_2,z_3) \\ + \\ \mu_2 G(z_2,z_3,z_3) \\ + \\ \mu_3 G(z_2,z_2,1) \end{cases} \tag{85}$$

from which we obtain

$$[\lambda(1-z_2)+\mu_2+\mu_3]\,G(z_2,z_2,z_3) = \begin{cases} \mu_2 G(z_2,z_3,z_3) \\ + \\ \mu_3 G(z_2,z_2,1)\ . \end{cases} \tag{86}$$

Here both the 'daughter' $G(z_2, z_2, 1)$ and the 'daughter' $G(z_2, z_3, z_3)$ are known and given by Eq. (56) and Eq. (84), respectively. We conclude that

$$G(z_2,z_2,z_3)$$

$$= \begin{cases} \frac{\mu_1\mu_2\mu_3 G(z_3,z_3,1)}{(\lambda(1-z_2)+\mu_1+\mu_3)(\lambda(1-z_2)+\mu_2+\mu_3)(\lambda(1-z_3)+\mu_3)} \\ + \\ \frac{\mu_2\mu_3 G(z_2,z_3,1)}{(\lambda(1-z_2)+\mu_1+\mu_3)(\lambda(1-z_2)+\mu_2+\mu_3)} \\ + \\ \frac{\mu_3 G(z_2,z_2,1)}{\lambda(1-z_2)+\mu_2+\mu_3}\ . \end{cases} \tag{87}$$

Substituting the expressions for $G(z_2, z_2, z_3)$ and $G(z_1, z_3, z_3)$ into Eq. (87) we obtain

$$[\lambda(1-z_1)+\mu]\,G(z_1,z_2,z_3)$$

$$= \begin{cases} \frac{\mu_1^2\mu_2\mu_3 G(z_3,z_3,1)}{(\lambda(1-z_2)+\mu_1+\mu_3)(\lambda(1-z_2)+\mu_2+\mu_3)(\lambda(1-z_3)+\mu_3)} \\ + \\ \frac{\mu_1\mu_2\mu_3 G(z_2,z_3,1)}{(\lambda(1-z_2)+\mu_1+\mu_3)(\lambda(1-z_2)+\mu_2+\mu_3)} \\ + \\ \frac{\mu_1\mu_3 G(z_2,z_2,1)}{\lambda(1-z_2)+\mu_2+\mu_3} \\ + \\ \frac{\mu_1\mu_2\mu_3 G(z_3,z_3,1)}{(\lambda(1-z_1)+\mu_1+\mu_3)(\lambda(1-z_3)+\mu_3)} \\ + \\ \frac{\mu_2\mu_3 G(z_1,z_3,1)}{\lambda(1-z_1)+\mu_1+\mu_3} \\ + \\ \mu_3 G(z_1,z_2,1) \end{cases} \tag{88}$$

Substituting the expressions for $G(z_3, z_3, 1)$, $G(z_2, z_3, 1)$, $G(z_2, z_2, 1)$, $G(z_1, z_3, 1)$ and $G(z_1, z_2, 1)$ into Eq. (88) we obtain the final expression for $G(z_1, z_2, z_3)$ given in Eq. (58).

6 Limit Laws

The analysis conducted in Chapter 5 concluded that the ASIP, despite its simple description, displays highly complex stochastic dynamics. An iterative scheme for the computation of the multidimensional probability generating function (PGF) of the ASIP's site occupancies at steady state was established. Yet, this PGF turns out to be analytically intractable even for small n — a fact that is vivid from the very rapid growth in complexity of the explicit PGF expressions for $n = 1, 2, 3$. Understanding the behavior of large ASIPs (i.e., ASIPs with large lattice size n) is therefore a challenge.

In this chapter we bypass the need for direct computation of occupancy PGFs and explore the stochastic limit laws of five key ASIP observables: (i) Traversal Time — the time it takes a particle to traverse the lattice; (ii) Overall Load — the total number of particles present in the lattice in steady state; (iii) Busy Period — the time elapsing from the instant a particle arrives at an empty lattice, till the first instant the lattice is empty once again; (iv) First Occupied Site — the index of the first non-empty site; (v) Draining Time — the time elapsing from the instant the arrival flow is blocked, in steady state, till the first instant the lattice is empty.

The stochastic limit laws of the above-mentioned observables are established in three different limiting regimes: (i) Heavy-Traffic regime — in which the particles' arrival rate λ tends to infinity; (ii) Large-System regime — in which the lattice size n tends to infinity; (iii) Balanced-System regime — in which the lattice size n tends to infinity, the gates' opening rates μ_k $(k = 1, \cdots, n)$ tend to infinity, and these limits are kept in balance. Our results hold for homogeneous and general (inhomogeneous) ASIPs alike.

The reminder of the chapter is organized as follows. We begin with a preliminary analysis of the ASIP's key observables (Section 6.1), analyze the asymptotic statistical behavior of homogeneous ASIPs (Section 6.2), compare our analytical results to simulations (Section 6.3), and then analyze the asymptotic statistical behavior of general ASIPs (Section 6.4).

6.1 Key Observables

In this section we analyze five key observables of the ASIP: Traversal Time, Overall Load, Busy Period, First Occupied Site, and Draining Time.

6.1.1 Traversal Time

The ASIP's *traversal time* T, first mentioned in Chapter 5, is the random time it takes a particle to traverse the lattice. Namely, T is the time elapsing from the instant a particle arrives at the first site $(k = 1)$, till the instant it leaves the last site $(k = n)$. Let us briefly recapitulate some basic facts regarding this random variable. Due to the memory-less property of the exponential distribution [59], the time elapsing from the arrival of a particle to site k (at an arbitrary time epoch), till the first opening of gate k thereafter, is exponentially distributed

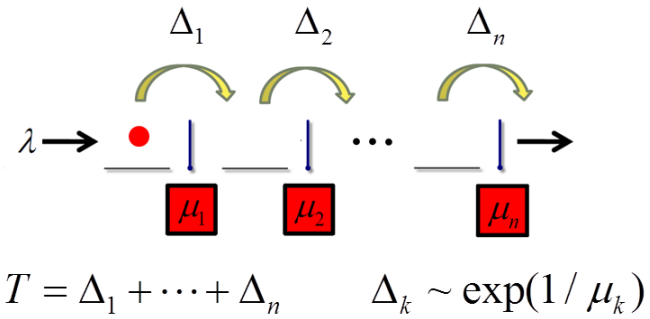

Figure 10: An illustration of the traversal time.

with mean $1/\mu_k$ — independently of the particle's arrival epoch to site k. A particle arriving to the lattice would thus wait an exponentially-distributed random time (with mean $1/\mu_1$) till moving from the first site to the second site, then wait an exponentially-distributed random time (with mean $1/\mu_2$) till moving from the second site to the third site, and so forth. Since the gate-openings are governed by independent Poisson processes we conclude that the traversal time T admits the stochastic representation

$$T = \sum_{k=1}^{n} \triangle_k \ , \tag{89}$$

where $\{\triangle_1, \cdots, \triangle_n\}$ is a sequence of independent and exponentially-distributed random times with corresponding means $\{1/\mu_1, \cdots, 1/\mu_n\}$.

Consequently, Eq. (89) straightforwardly implies that the mean and the Laplace transform of the traversal time T are given, respectively, by

$$\mathbf{E}\left[T\right] = \sum_{k=1}^{n} \frac{1}{\mu_k} \tag{90}$$

and

$$\mathbf{E}\left[\exp\left(-\theta T\right)\right] = \prod_{k=1}^{n} \frac{\mu_k}{\mu_k + \theta} \tag{91}$$

($\theta \geq 0$). The traversal time is illustrated schematically in Figure 10.

6.1.2 Overall Load

Consider the ASIP in *steady state*, and let X_k denote the number of particles present in site k ($k = 1, \cdots, n$). The ASIP's *overall load* L, first mentioned in Chapter 4, is the *total number of particles* present in the lattice in steady state:

$$L = \sum_{k=1}^{n} X_k \ . \tag{92}$$

Let us briefly recapitulate some basic facts regarding this random variable. The analysis presented in Section 5.7 showed that the mean overall load is given by

$$\mathbf{E}[L] = \sum_{k=1}^{n} \frac{\lambda}{\mu_k} \tag{93}$$

Eq. (93), in conjunction with Eq. (90), implies that $\mathbf{E}[L] = \lambda \mathbf{E}[T]$. Namely, the mean overall load $\mathbf{E}[L]$ equals the product of the inflow rate λ and the mean traversal time $\mathbf{E}[T]$ — the mean sojourn time of an arbitrary particle in the lattice. Equation (93) is the ASIP's version of the well known *Little's law* in Queueing Theory [9].

The analysis presented in Section 5.7 further established that the probability generating function of the overall load L is given by

$$\mathbf{E}\left[z^{L}\right] = \prod_{k=1}^{n} \frac{\mu_k}{\mu_k + \lambda(1-z)} \tag{94}$$

($|z| \leq 1$). Eq. (94) has several important implications. First, it implies that the overall load L of a *single-site* ASIP ($n = 1$) follows a *geometric distribution*: For $n = 1$ the probability generating function of Eq. (94) yields the probability distribution $\Pr(L = j) = (1 - p_1)^j p_1$ ($j = 0, 1, 2 \cdots$), where $p_1 = \mu_1 / (\mu_1 + \lambda)$. Second, the *product-form* structure of Eq. (94) implies that the overall load L admits the stochastic representation

$$L = \sum_{k=1}^{n} G_k \ , \tag{95}$$

where $\{G_1, \cdots, G_n\}$ is a sequence of independent geometrically-distributed random variables: $\Pr(G_k = j) = (1-p_k)^j p_k$ ($j = 0, 1, 2 \cdots$), where $p_k = \mu_k / (\mu_k + \lambda)$. The overall load L is hence equal, in law, to the sum of the overall loads of n *independent single-site* ASIPs with respective parameters $(\lambda, \mu_1), \cdots, (\lambda, \mu_n)$. Third, Eq. (94) implies the following distributional form of the aforementioned *ASIP Little's law*: $L = \Pi_0(T)$, the equality being in law. Namely, the number of particles arriving to the lattice, $\Pi_0(T)$, during a traversal time T is equal, in law, to the ASIP's overall load L. The proof of the distributional Little's law is given in the Appendix to this chapter. Fourth, setting $z = 0$ in Eq. (94) implies that the probability that the lattice is empty is given by

$$\Pr(L = 0) = \prod_{k=1}^{n} \frac{\mu_k}{\mu_k + \lambda} \ . \tag{96}$$

The overall load is illustrated schematically in Figure 11.

6.1.3 Busy Period

The ASIP's *busy period* B is the random duration of time in which the lattice is continuously non-empty. Namely, B is the time elapsing from the instant a

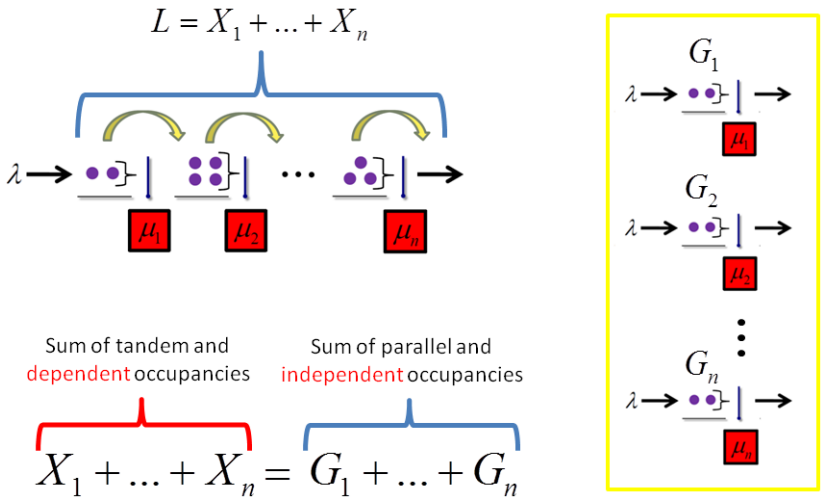

Figure 11: An illustration of the overall load.

particle arrives at an empty lattice, till the first instant thereafter the lattice is empty once again. The busy period is a key variable in queueing theory, as every queueing system continuously alternates between random busy and idle periods [68, 69, 70].

Consider the two following scenarios: (i) a particle arrives at an empty lattice and traverses it before a second particle arrives; (ii) a particle arrives at an empty lattice and a second particle arrives before the first particle traversed the lattice. Let T denote the traversal time of the first particle, and let $\triangle_0$ denote the time elapsing between the arrival epochs of the two particles. Clearly, the random variables T and $\triangle_0$ are independent.

The first scenario is the event $\{T < \triangle_0\}$, and in this scenario the busy period equals the traversal time: $B = T$. The second scenario is the event $\{\triangle_0 \leq T\}$, and in this scenario the busy period equals the inter-arrival time $\triangle_0$ plus an additional and independent random time whose distribution is equal in law to that of a busy period: $B = \triangle_0 + B'$, where B' is an IID copy of B which is independent of T and $\triangle_0$. Thus, we obtain that the busy period B satisfies the following stochastic regeneration formula

$$B = \begin{cases} T & \text{if } T < \triangle_0 \ , \\ \\ \triangle_0 + B' & \text{if } \triangle_0 \leq T \ . \end{cases} \tag{97}$$

Consequently, Eq. (97) implies that the mean and the Laplace transform of the busy period B are given, respectively, by

$$\mathbf{E}\left[B\right] = \frac{1}{\lambda}\left(\prod_{k=1}^{n}\left[1 + \frac{\lambda}{\mu_k}\right] - 1\right) \tag{98}$$

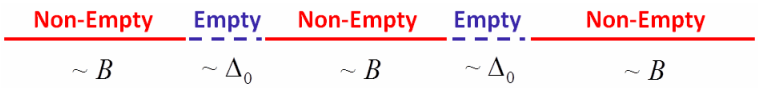

Figure 12: An illustration of the renewal argument (in the context of the mean busy period).

and

$$\mathbf{E}\left[\exp\left(-\theta B\right)\right] = \frac{\lambda+\theta}{\lambda+\theta\prod_{k=1}^{n}\left[1+\frac{\lambda+\theta}{\mu_k}\right]} \tag{99}$$

$(\theta \geq 0)$.

The derivations of Eqs. (98) and (99) are given in the Appendix to this chapter. Eq. (98) can also be obtained via a renewal argument which we now describe.

Note that the lattice alternates between empty and non-empty periods. The empty periods are IID copies of the generic inter-arrival period $\triangle_0$, the non-empty periods are IID copies of the generic busy period B, and these alternating periods are mutually independent. Renewal theory implies that — over infinitely large time windows — the fraction of time the lattice is empty is given by the ratio [71]: $\mathbf{E}\left[\triangle_0\right] / \left(\mathbf{E}\left[\triangle_0\right] + \mathbf{E}\left[B\right]\right)$. On the other hand, the fraction of time the lattice is empty equals the probability, in steady state, of a zero overall load: $\Pr\left(L=0\right)$. Thus, we obtain that $\Pr\left(L=0\right) = \mathbf{E}\left[\triangle_0\right] / \left(\mathbf{E}\left[\triangle_0\right] + \mathbf{E}\left[B\right]\right)$. Now, since $\mathbf{E}\left[\triangle_0\right] = 1/\lambda$, and since $\Pr\left(L=0\right)$ is given by Eq. (96), we can extract the mean busy period $\mathbf{E}\left[B\right]$. Doing so indeed yields Eq. (98). The renewal argument is illustrated schematically in Figure 12.

6.1.4 The First Occupied Site

Consider the ASIP in *steady state*, and let I denote the index of the first occupied site: $I = \min\{k | X_k > 0\}$. If all the sites are empty then we set $I = \infty$ by convention. Clearly

$$\begin{aligned} &\Pr(I=k) = \Pr(X_1=0,\cdots,X_{k-1}=0,X_k>0) \\ &= \Pr(X_1=0,\cdots,X_{k-1}=0) - \Pr(X_1=0,\cdots,X_k=0) \end{aligned} \tag{100}$$

$(1 \leq k \leq n)$, and

$$\Pr\left(I=\infty\right) = \Pr\left(X_1=0,\cdots,X_n=0\right) \ . \tag{101}$$

Since the event $\{X_1=0,\cdots,X_n=0\}$ is equivalent to the event$\{L=0\}$, the probability appearing on the right-hand-side of Eq. (101) is given by Eq. (96). In order to compute the probabilities appearing on the right hand side of Eq. (100) we remind the reader of the embedding property that was derived in the end of Section 5.5.

Consider two ASIPs: ASIP A with n' gates and parameters $\{\lambda, \mu_1, \cdots, \mu_{n'}\}$, and ASIP B with n gates and parameters $\{\lambda, \mu_1, \cdots, \mu_n\}$, where $n' \leq n$. The embedding property asserts that the steady state distribution of ASIP A coincides with the steady state distribution of the first n' sites of ASIP B. An intuitive understanding of the embedding phenomenon follows from the fact that in an ASIP model with n gates the operation of the first n' gates is independent of whatever happens in the following gates $\{n'+1, \cdots, n\}$. In other words, an observation of the first n' gates in an ASIP with n gates is indistinguishable from an observation of an ASIP with n' gates (and the same parameters).

Due to the embedding phenomenon Eq. (96) implies that $\Pr(X_1 = 0, \cdots, X_{n'} = 0) = \prod_{k=1}^{n'} \frac{\mu_k}{\mu_k + \lambda}$ $(1 \leq n' \leq n)$. Substituting these probabilities into Eqs. (100) and (101) we obtain that the distribution of the first occupied site I is given by

$$\begin{cases} \Pr(I = k) = \frac{\lambda}{\mu_k + \lambda} \prod_{j=1}^{k-1} \frac{\mu_j}{\mu_j + \lambda} \quad (1 \leq k \leq n)\,, \\ \\ \Pr(I = \infty) = \prod_{j=1}^{n} \frac{\mu_j}{\mu_j + \lambda}\,. \end{cases} \tag{102}$$

6.1.5 Draining Time

Consider the ASIP in *steady state*, and assume that — starting at an arbitrary time epoch — we *block* the inflow of newcoming particles to the lattice. The ASIP's *draining time* D is the duration of time it takes the lattice to clear. Namely, D is the random time elapsing from the blocking epoch till the first instant the lattice is empty.

Consider the index I of the first occupied site at the blocking epoch. If $I = k$ $(k = 1, \cdots, n)$ then the draining time equals the traversal time of the gates $\{k, k+1, \cdots, n\}$. Consequently — analogous to the derivation of Eq. (89) — we obtain that if $I = k$ then $D = \sum_{j=k}^{n} \triangle_j$, where $\{\triangle_k, \cdots, \triangle_n\}$ is a sequence of independent and exponentially-distributed random times with corresponding means $\{1/\mu_k, \cdots, 1/\mu_n\}$. On the other hand, if $I = \infty$ then the lattice is empty and hence $D = 0$. Thus, we obtain that the draining time admits the stochastic representation

$$D = \begin{cases} \sum_{j=k}^{n} \triangle_j & \text{if } I = k \quad (1 \leq k \leq n)\,, \\ \\ 0 & \text{if } I = \infty\,, \end{cases} \tag{103}$$

where the exponentially-distributed random variables $\{\triangle_1, \cdots, \triangle_n\}$ are independent of the first occupied site I.

Since we blocked the inflow to an ASIP in *steady state* the distribution of the first occupied site I is given by Eq. (102). Consequently, combining together Eqs. (102) and (103) we obtain that the mean and the Laplace transform of the

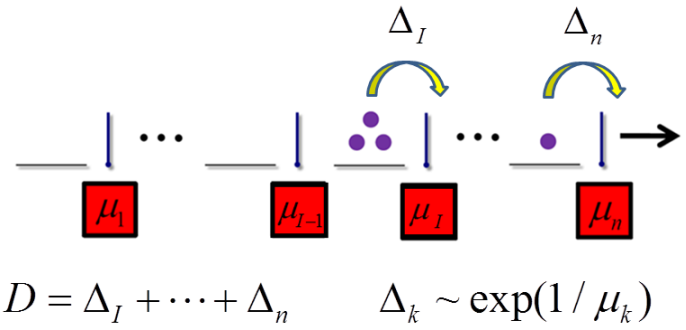

Figure 13: An illustration of the draining time.

draining time D are given, respectively, by

$$\mathbf{E}[D] = \sum_{k=1}^{n} \frac{\lambda}{\mu_k + \lambda} \left(\prod_{j=1}^{k-1} \frac{\mu_j}{\mu_j + \lambda} \right) \left(\sum_{j=k}^{n} \frac{1}{\mu_j} \right) \tag{104}$$

and

$$\begin{aligned} \mathbf{E}[\exp(-\theta D)] &= \prod_{j=1}^{n} \tfrac{\mu_j}{\mu_j+\lambda} \\ &+ \sum_{k=1}^{n} \tfrac{\lambda}{\mu_k+\lambda} \left(\prod_{j=1}^{k-1} \tfrac{\mu_j}{\mu_j+\lambda} \right) \left(\prod_{j=k}^{n} \tfrac{\mu_j}{\mu_j+\theta} \right) \end{aligned} \tag{105}$$

($\theta \geq 0$). The derivations of Eqs. (104) and (105) are given in the Appendix to this chapter. The draining time is illustrated schematically in Figure 13.

6.2 Asymptotic Analysis: The Homogeneous Case

In this section we consider homogeneous ASIPs, and set $\mu_1 = \cdots = \mu_n = 1/m$. Namely, m is the mean sojourn time of an arbitrary particle in a single site. In what follows we shall establish stochastic limit laws for the ASIP's key observables presented in the previous section: Traversal Time, Overall Load, Busy Period, First Occupied Site, and Draining Time. The stochastic limit laws shall be established for the three following limiting regimes:

- The *Heavy-Traffic regime* — in which the inflow rate tends to infinity: $\lambda \to \infty$.
- The *Large-System regime* — in which the lattice size tends to infinity: $n \to \infty$.
- The *Balanced-System regime* — in which the lattice size tends to infinity, the particles' mean sojourn time at a site tends to zero, and the product of these parameters tends to a positive limit: $n \to \infty$, $m \to 0$, and $nm \to \tau \in (0, \infty)$.

Throughout this section we denote by Z a Gauss-distributed random variable with zero mean and unit variance ("Standard Normal"), by $\mathcal{E}$ an exponentially distributed random variable with unit mean ("Standard Exponential"), and by Γ_n an Erlang distributed random variable with n degrees of freedom. Namely, Γ_n is the sum of n IID copies of the random variable $\mathcal{E}$. In what follows the sign "$\approx$" will denote asymptotic equivalence in law.

6.2.1 Heavy Traffic

The heavy-traffic regime considers ASIPs in which the inflow rate is increased to infinity: $\lambda \to \infty$. The ASIP stochastic limit laws — under the heavy-traffic regime — are as follows:

Traversal Time. As is clear from Eq. (91), the inflow rate does not affect the traversal time T. On the other hand, in the homogeneous ASIP the traversal time T is the sum of n IID exponential random variables each with mean m. Hence, the traversal time T admits the stochastic representation

$$T = m\Gamma_n \tag{106}$$

(recall that a random variable $\mathcal{E}$ is exponentially distributed with unit mean if and only if the random variable $m\mathcal{E}$ is exponentially distributed with mean m).

Overall Load. Increasing the inflow rate λ is expected to result in an increase of the overall load L. And indeed, Eq. (93) implies that the mean of the overall load L scales linearly with λ. Consequently, we normalize the overall load L by the dimensionless term λm and analyze the stochastic limit of the normalized overall load $L/(\lambda m)$ (as $\lambda \to \infty$). Setting $z = \exp(-\theta/(\lambda m))$ in Eq. (94) we obtain the limit

$$\lim_{\lambda\to\infty} \mathbf{E}\left[\exp\left(-\theta\frac{L}{\lambda m}\right)\right] = \left(\frac{1}{1+\theta}\right)^n \tag{107}$$

($\theta \geq 0$). Since the right-hand-side of Eq. (107) is the Laplace transform of the Erlang distribution with n degrees of freedom, we obtain that, as $\lambda \to \infty$, the overall load L admits the stochastic approximation

$$L \approx \lambda \cdot m\Gamma_n \tag{108}$$

($\lambda \to \infty$).

Busy Period. As in the case of the overall load, increasing the inflow rate λ is expected to result in an increase of the duration of the busy period B. And indeed, Eq. (98) implies that the mean of the busy period B scales like λ^{n-1}. Consequently, we normalize the busy period B by the dimensionless term $(m\lambda)^{n-1}$ and analyze the stochastic limit of the normalized busy period $B/(m\lambda)^{n-1}$ (as $\lambda \to \infty$). Using Eq. (99) we obtain the limit

$$\lim_{\lambda\to\infty} \mathbf{E}\left[\exp\left(-\theta\frac{B}{(m\lambda)^{n-1}}\right)\right] = \frac{1}{1+m\theta} \tag{109}$$

($\theta \geq 0$). Since the right-hand-side of Eq. (109) is the Laplace transform of the exponential distribution with mean m, we obtain that the busy period B admits the stochastic approximation

$$B \approx \lambda^{n-1} \cdot m^n \mathcal{E} \tag{110}$$

($\lambda \to \infty$).

First Occupied Site. Increasing the inflow rate λ is expected to increase to one the probability of finding the first site occupied. And indeed, Eq. (102) yields the limit

$$\lim_{\lambda \to \infty} \Pr(I = 1) = 1 \ . \tag{111}$$

Draining Time. Equation (111) implies that for large λ the first occupied site is effectively the first site. Consequently, for large λ the draining time D should coincide with the traversal time T. And indeed, taking the limit $\lambda \to \infty$ in Eq. (105) confirms this conjecture and yields the stochastic approximation

$$D \approx m\Gamma_n \tag{112}$$

($\lambda \to \infty$).

6.2.2 Large Systems

The large-system regime considers ASIPs in which the lattice size increases to infinity: $n \to \infty$. The ASIP stochastic limit laws — under the large-system regime — are as follows:

Traversal Time. In the homogeneous ASIP the traversal time T is a sum of n IID exponential random variables — each with mean m and variance m^2. Consequently, the Central Limit Theorem [59] implies that the traversal time T admits the Gaussian stochastic approximation

$$T \approx n \cdot m + \sqrt{n} \cdot mZ \tag{113}$$

($n \to \infty$).

Overall Load. Equation (94) asserts that in the homogeneous ASIP the overall load L is a sum of n IID geometric random variables — each with mean λm and variance $\lambda m + (\lambda m)^2$. Consequently, the Central Limit Theorem [59] implies that the overall load L admits the Gaussian stochastic approximation

$$L \approx n \cdot \lambda m + \sqrt{n} \cdot \sqrt{\lambda m + (\lambda m)^2}\, Z \tag{114}$$

($n \to \infty$).

Busy Period. Increasing the lattice size n is expected to result in an increase of the busy period. Indeed, Eq. (98) implies that for large n the mean of the busy period scales like $(1 + \lambda m)^n$. Consequently, we normalize the busy period by the dimensionless term $(1 + \lambda m)^n$ and analyze the stochastic limit

of the normalized busy period $B/(1+\lambda m)^n$ (as $n \to \infty$). Using Eq. (99) we obtain the limit

$$\lim_{n\to\infty} \mathbf{E}\left[\exp\left(-\theta \frac{B}{(1+\lambda m)^n}\right)\right] = \frac{\lambda}{\lambda+\theta} \tag{115}$$

($\theta \geq 0$). Since the right-hand-side of Eq. (115) is the Laplace transform of the exponential distribution with mean $1/\lambda$, we obtain that the busy period B admits the stochastic approximation

$$B \approx (1+\lambda m)^n \cdot \frac{1}{\lambda}\mathcal{E} \tag{116}$$

($n \to \infty$).

First Occupied Site. Taking the limit $n \to \infty$ in Eq. (102) yields the geometric distribution

$$\lim_{n\to\infty} \Pr(I = k) = q\,(1-q)^{k-1} \tag{117}$$

($k = 1, 2, 3, \cdots$), where $q = \lambda m/(1+\lambda m)$.

Draining Time. Equation (117) implies that — in the limit $n \to \infty$ — the First Occupied Site is geometrically distributed with mean $1/q$ and that $\mathbf{E}\left[\sum_{k=1}^{I-1} \triangle_k\right]$ is a finite constant that does not depend on n. Consequently — since the number of sites tends to infinity ($n \to \infty$) — the draining time D effectively equals the traversal time T. Combining this observation together with Eq. (113) implies that the draining time D admits the Gaussian stochastic approximation

$$D \approx n \cdot m + \sqrt{n} \cdot mZ \tag{118}$$

($n \to \infty$). A rigorous proof of this result is given in the Appendix to this chapter.

6.2.3 Balanced Systems

The balanced-system regime considers ASIPs in which the number of sites increases to infinity ($n \to \infty$) and the mean sojourn time at a site decrease to zero ($m \to 0$) — while their product tends to a positive limit: $nm \to \tau \in (0, \infty)$. Namely, in this regime the large number of sites is balanced by rapid gate-opening rates. With no loss of generality we henceforth set $m = \tau/n$ and consider the limit $n \to \infty$. The ASIP stochastic limit laws — under the balanced-system regime — are as follows:

Traversal Time. Setting $m = \tau/n$ into Eq. (91) we obtain the limit

$$\lim_{n\to\infty} \mathbf{E}\left[\exp\left(-\theta T\right)\right] = \exp\left(-\theta\tau\right) \tag{119}$$

($\theta \geq 0$). The right-hand-side of Eq. (119) is the Laplace transform of a degenerate random variable which admits the value τ with probability one. Thus, the traversal time converges in law to the *deterministic* value τ.

Overall Load. Setting $m = \tau/n$ into Eq. (94) we obtain the limit

$$\lim_{n\to\infty} \mathbf{E}\left[z^L\right] = \exp\left(-\tau\lambda\left(1-z\right)\right) \tag{120}$$

($|z| \leq 1$). The right-hand-side of Eq. (120) is the probability generating function of the Poisson distribution with mean $\tau\lambda$. Thus, the overall load converges in law to a Poisson random variable with mean $\tau\lambda$.

Busy Period. Setting $m = \tau/n$ into Eq. (99) we obtain the limit

$$\lim_{n\to\infty} \mathbf{E}\left[\exp\left(-\theta B\right)\right] = \frac{\lambda+\theta}{\lambda+\theta\exp\left(\tau\left(\lambda+\theta\right)\right)} \tag{121}$$

($\theta \geq 0$). The right-hand-side of Eq. (121) can be derived from the stochastic regeneration formula of Eq. (97) by setting $T = \tau$ — see Appendix for details. We note that (in the limit $n \to \infty$) the busy period B is equal to τ with probability $\exp\left(-\lambda\tau\right)$, and is larger than τ otherwise. Indeed, $B = \tau$ if and only if there are no particle arrivals during the traversal time τ — an event which takes place with probability $\exp\left(-\lambda\tau\right)$. Also, Eq. (121) implies that (in the limit $n \to \infty$) the mean of the busy period is given by $\left(\exp\left(\lambda\tau\right)-1\right)/\lambda$.

First Occupied Site. Consider the scaled first occupied site $\hat{I} = I/n$. Setting $m = \tau/n$ into Eq. (102) we obtain the limits

$$\begin{cases} \lim_{n\to\infty} \Pr\left(\hat{I} \leq x\right) = 1 - \exp\left(-\tau\lambda x\right) \\ \\ \lim_{n\to\infty} \Pr(\hat{I} = \infty) = \exp\left(-\lambda\tau\right) \end{cases} \tag{122}$$

($0 \leq x \leq 1$); recall that the event $\{\hat{I} = \infty\}$ represents the (steady-state) scenario in which all sites are empty. Equation (122) implies that the scaled first occupied site $\hat{I}$ converges, in law, to a limit which has the density of an exponential random variable, with mean $\lambda\tau$, on the unit interval and an atom with probability $\exp\left(-\lambda\tau\right)$ at infinity. The derivation of Eq. (122) is given in the Appendix to this chapter. Hence the scaled first occupied site $\hat{I}$ admits the asymptotic stochastic approximation

$$\hat{I} \approx \begin{cases} \mathcal{E}/\left(\lambda\tau\right) & \text{if } \mathcal{E} \leq \lambda\tau \\ \\ \infty & \text{if } \mathcal{E} > \lambda\tau \end{cases} \tag{123}$$

($n \to \infty$).

Draining Time. Setting $m = \tau/n$ into Eq. (105) we obtain the limit

$$\lim_{n\to\infty} \mathbf{E}\left[\exp\left(-\theta D\right)\right] = \frac{\theta\exp\left(-\lambda\tau\right) - \lambda\exp\left(-\theta\tau\right)}{\theta-\lambda} \tag{124}$$

($\theta \geq 0$). The derivation of Eq. (124) is given in the Appendix to this chapter. Equation (124) implies that an asymptotic stochastic approximation for the draining time is given by an amalgamation of a probability density function

$$f_D(x) = \begin{cases} \exp\left(-\lambda\left(\tau-x\right)\right) & 0 < x \leq \tau \\ \\ 0 & \textit{otherwise} \end{cases} \tag{125}$$

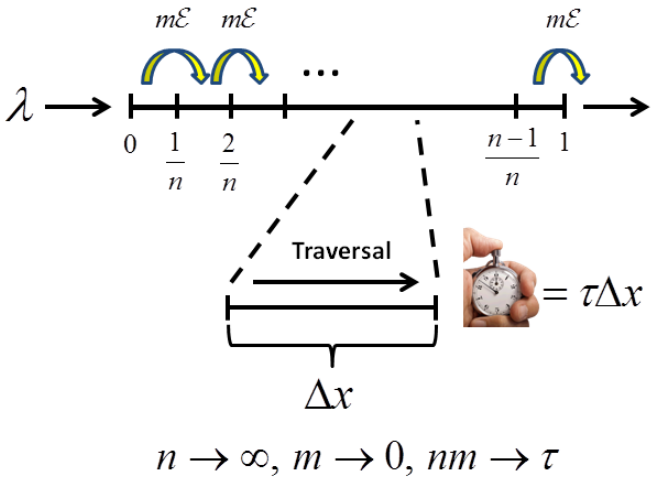

Figure 14: An illustration of the homogeneous ASIP in the balanced-system limiting regime.

and a probability mass, $\exp(-\lambda\tau)$, at zero. The validity of this approximation is easily verified by taking Laplace transform and recovering the right-hand-side of Eq. (124).

Mapping the ASIP's lattice to the unit interval — positioning site k at the point $x = k/n$ — the balanced-system limiting regime is illustrated schematically in Figure 14.

6.2.4 The M/D/∞ Queue

The balanced-system limiting regime may be better understood in light of the mapping between homogeneous ASIPs in this regime and the $M/D/\infty$ queue [72], which we describe herein.

In queueing theory, the $M/D/\infty$ queue represents a system consisting of an infinite number of servers (or infinite "broadband"), to which particles ("jobs" in the queueing jargon) arrive following a Poisson process with rate λ. Each particle, upon its arrival, is immediately attended by one of the available servers; upon service completion a served particle leaves the system. Service times are deterministic and of common length τ, and the particles are served independently. Note that the common deterministic service times assure that particles will leave the system *exactly* τ units of time after their respective arrival epochs, and will do so in a *First In First Out* (FIFO) manner.

From the particles' perspective the ASIP — in the balanced-system limiting regime — is identical to the $M/D/\infty$ queue. Indeed, particles arrive to the lattice following a Poisson process with rate λ. Each particle, upon its arrival to the lattice, starts traversing it. In the balanced-system limiting regime the particles' traversal times are deterministic and of common length τ; upon "traversal

completion" the particles leave the lattice. Here again, the common deterministic traversal times assure that particles will leave the lattice *exactly* τ units of time after their respective arrival epochs, and will do so separately — i.e., one by one — and in a FIFO manner. We emphasize that in the balanced-system limiting regime the particles remain separated and thus do *not* coalesce into clusters.

Thus, the ASIP in the balanced-system regime — with deterministic traversal time τ — is identical to a $M/D/\infty$ queue with deterministic service time τ. This observation gives rise to additional analogies between ASIPs in the limiting balanced-system regime and the $M/D/\infty$ queue. (i) The ASIP's overall load is equal in law to the $M/D/\infty$ queue size — the total number of jobs present (i.e., being served) in the $M/D/\infty$ queue in steady state, and is distributed according to the Poisson distribution with mean $\tau\lambda$ [72]. (ii) The ASIP's busy period is equal in law to the busy period of the $M/D/\infty$ queue. Indeed, the busy period in the $M/D/\infty$ queue follows the stochastic regeneration formula

$$B = \begin{cases} \tau & \text{if } \tau < \triangle_0 \ , \\ \\ \triangle_0 + B' & \text{if } \triangle_0 \leq \tau \ , \end{cases} \tag{126}$$

in which τ is the service time, $\triangle_0$ is the exponential time elapsing between the arrival epochs of two jobs and B' is an IID copy of the busy period B. Note that Eq. (126) is a special case of Eq. (97) — in which the general traversal time T is replaced by the deterministic traversal time τ. (iii) The ASIP's draining time is equal in law to the residual service-completion time of the newest job in the $M/D/\infty$ queue size. Indeed, the cumulative distribution function of the residual service-completion time, T_{res}, in the $M/D/\infty$ queue is equal to the probability that the inter-arrival time between jobs will exceed the value $\tau - t$ and is hence given by:

$$\Pr\left(T_{res} \leq t\right) = \exp\left(-\lambda(\tau - t)\right) \tag{127}$$

$(0 \leq t \leq \tau)$. One can easily verify that the Laplace transform of T_{res} coincides with the right hand side of Eq. (124). The mapping between the balanced-system limit of the ASIP and the $M/D/\infty$ queue is illustrated schematically in Figure 15.

6.3 Comparison With Simulations

In this section we compare the limit laws obtained above with numerical simulations. Our main aim is to visually demonstrate convergence, and to illustrate how general thumb rules regarding the applicability of the limiting distributions as useful approximations can be attained. Throughout the section we use the following set of parameters:

- *Heavy-Traffic regime:* $m = 1$ and $n = 10$.
- *Large-System regime:* $m = 1$ and $\lambda = 1$.

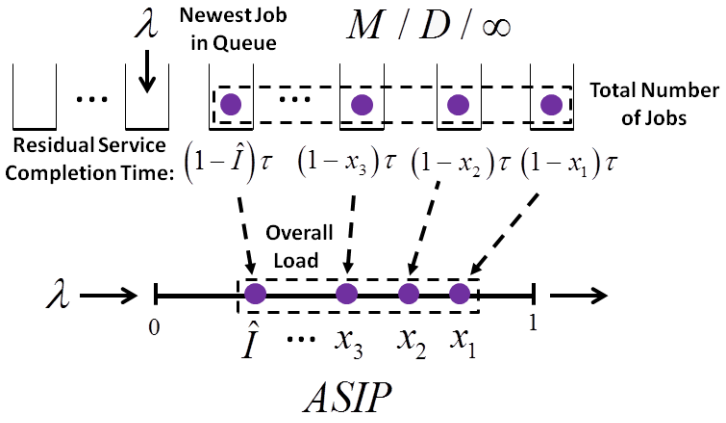

Figure 15: The ASIP's balanced-system limiting regime and the mapping to the $M/D/\infty$ queue. Each job in the $M/D/\infty$ queue is mapped to a particle on the unit interval according to its residual service time. Particles move at a constant velocity $1/\tau$, the residual service time in the $M/D/\infty$ queue is hence τ times the distance from the right edge of the unit interval.

- *Balanced-System regime:* $\lambda = 1$ and $nm = 1$.

We note that without loss of generality either λ or m can always be set to unity.

Traversal Time. Recalling that the traversal time does not depend on the particles' arrival rate λ, we examine the asymptotic behavior of this observable in the Large-System and Balanced-System regimes only.

In the top panel of Figure 16 we plot histograms of the standardized traversal time, $(T-n)/\sqrt{n}$, in the Large-System regime. As predicted by Eq. (113), with $m = 1$, convergence to the standard Gaussian distribution is visible as the bell shape curve gradually takes the place of the positively skewed, Erlang like, distribution that characterizes the standardized traversal time for small n.

In the bottom panel of Figure 16 we plot histograms of the traversal time, T, in the Balanced-System regime. As n increases, histograms become sharply peaked around unity reflecting the convergence to a deterministic random variable, as predicted by Eq. (119). To that end we note that in the Balanced-System regime the standard deviation of the traversal time is of order $O(n^{-1/2})$.

Overall Load. In the top panel of Figure 17 we plot cumulative distribution functions of the load in the Heavy-traffic regime. As predicted by Eq. (107), with $n = 10$, convergence to the Erlang distribution with ten degrees of freedom is clearly visible as simulated curves virtually collapse onto the Erlang curve for $\lambda \geq 25$. Convergence is also visible when plotting histograms as we do in the inset. Doing so, one should keep in mind that the load is a discrete random variable for which there is no proper probability density. This fact creates a somewhat deceiving impression regarding convergence, as the onset of "density like" histograms strongly depends on the preselected bin widths which in turn

affect bar heights in the histogram.

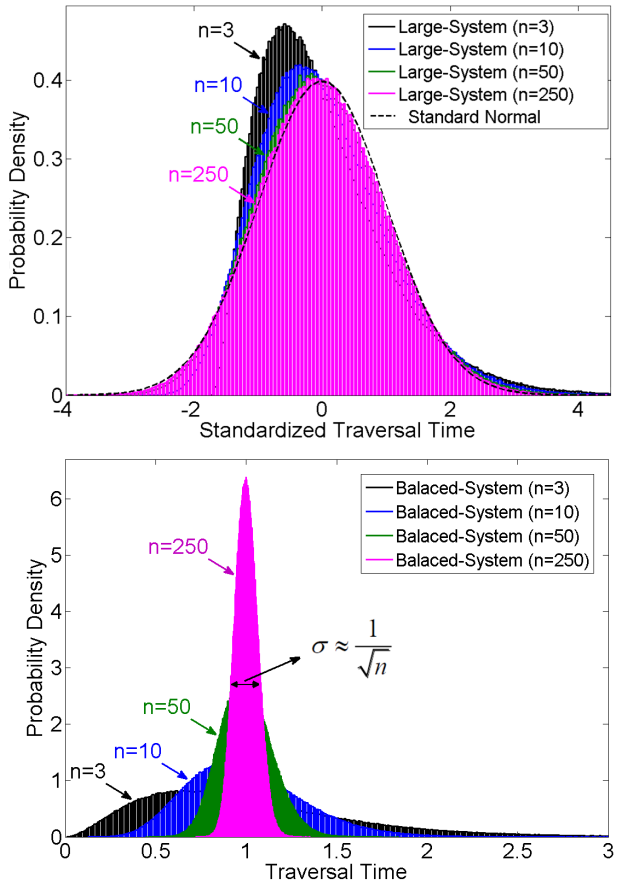

Figure 16: Asymptotic behavior of the traversal time in the Large-System (top panel) and Balanced-System (bottom panel) regimes.

In the middle panel of Figure 17 we plot cumulative probability functions of the standardized load, $(L-n)/\sqrt{2n}$, in the Large-System regime. As predicted by Eq. (114), with $\lambda = m = 1$, convergence to the standard Gaussian distribution is clearly visible as the simulated curves closely follow the standard Gaussian curve even for moderate values of n. Convergence is also visible when plotting histograms as we do in the inset.

In the bottom panel of Figure 17 we plot the ratio between the probability that the overall load in the system is k ($k = 0, 1, 2, .., 6$) and the limiting probability of this event in the Balanced-System regime, for several different values of n. In this type of plot, every deviation of the ratio from unity can be interpreted as a deviation from the Poissonian limit law given by Eq. (120). Values that are smaller/larger than unity mean that the observed probability

will be over/under estimated by the Poissonian approximation. As n increases convergence to the Poissonian limit (bars of unit height) is clearly visible and it can be considered a fair approximation even for moderate values of n. We note that under the chosen set of parameters, the total error made by neglecting $k > 6$ terms is given by the probability tail $Pr(L > 6) \cong 8.32 \cdot 10^{-5}$.

Busy Period. In the top panel of Figure 18 we plot probability density curves of the normalized busy period, B/λ^9, in the Heavy-traffic regime. As predicted by Eq. (109), with $m = 1$ and $n = 10$, convergence to the exponential distribution with unit mean ("Standard Exponential") is clearly visible. Under this choice of parameters the exponential approximation can be considered very good for $\lambda \geq 250$.

In the middle panel of Figure 18 we plot probability density curves of the normalized busy period, $B/2^n$, in the Large-System regime. As predicted by Eq. (115), with $\lambda = m = 1$, rapid convergence to the exponential distribution with unit mean ("Standard Exponential") is clearly visible. Under this choice of parameters the exponential approximation can be considered very good even for relatively small ($n \geq 10$) values of n.

In the bottom panel of Figure 18 we plot cumulative distribution functions of the busy period in the Balanced-System regime. Convergence to the asymptotic cumulative distribution function predicted by numerical inversion of the Laplace transform given in Eq. (121), with $\lambda = \tau = 1$, is clearly visible. However, convergence seems slower than in the Heavy-Traffic and Large-System regimes and the asymptotic distribution can be considered a fair approximation only for relatively large values of n ($n \geq 1250$). This slow convergence is due to the discontinuity (at $B = 1$) of the cumulative distribution function of the limiting busy-period.

First Occupied Site and Draining Time. In the top panel of Figure 19 we plot histograms of the draining time in the Heavy-traffic regime. As predicted by Eq. (112), with $n = 10$, convergence to the Erlang distribution with ten degrees of freedom is clearly visible and the Erlang approximation seems excellent even for relatively small particles' arrival rate ($\lambda = 3$). In the inset we plot the probability of finding the first site occupied as a function of λ. Under the chosen set of parameters this probability is given by $\Pr(I = 1) = \lambda/(1 + \lambda)$. As delineated by Eq. (111) this probability rapidly approaches unity as λ increases.

In the middle panel of Figure 19 we plot histograms of the standardized draining time, $(D - n)/\sqrt{n}$, in the Large-System regime. As predicted by Eq. (118), with $m = 1$, convergence to the standard Gaussian distribution is clearly visible and the Gaussian approximation seems fair for $n \geq 250$. Equation (102) asserts that for homogeneous ASIPs the first occupied site follows a *truncated* geometric distribution for which $\Pr(I = k) = q(1 - q)^{k-1}$ $(k = 1, 2, 3, \cdots, n)$, where $q = \lambda m/(1 + \lambda m)$ and $\Pr(I = \infty) = (1 + \lambda m)^n$. Hence, the probability that $I = \infty$, i.e. all sites are empty, can also be understood as the total error made in approximating $\Pr(I = k)$ $(k = 1, 2, 3, \cdots)$ by the geometric limit, given in Eq. (117). Under the chosen set of parameters $\Pr(I = \infty) = 2^{-n}$, and the total error made by making use of the geometric approximation rapidly decays to zero as is clearly illustrated in the inset.

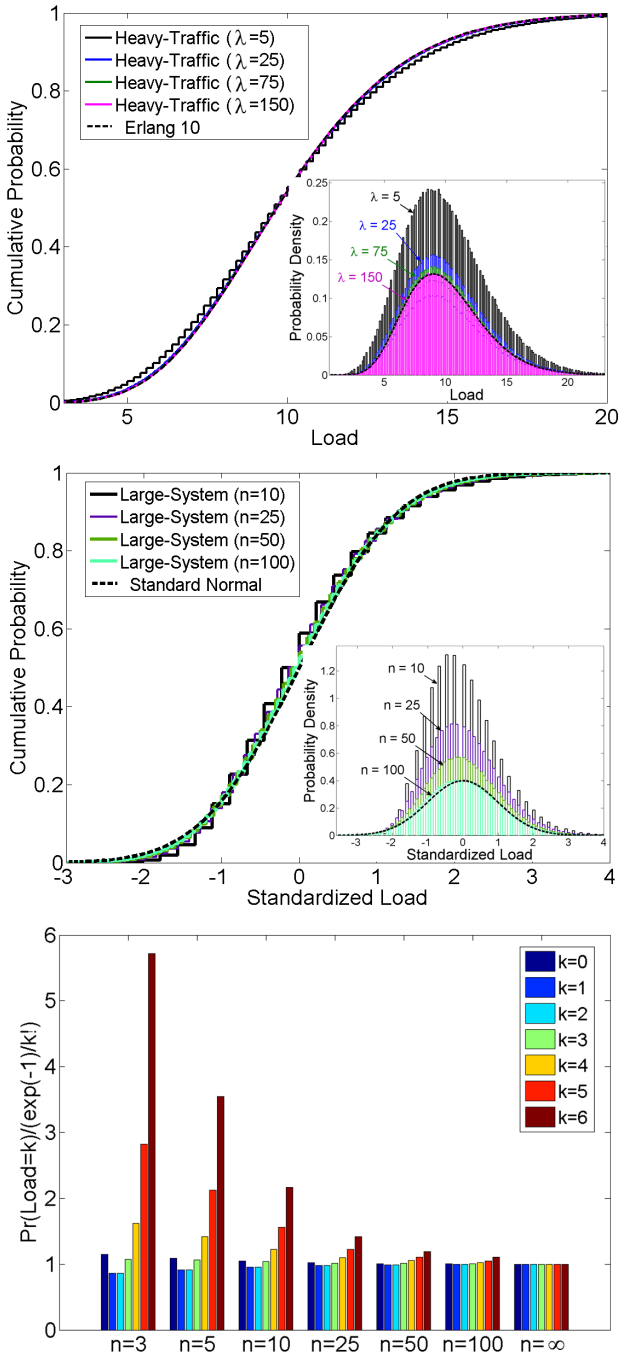

Figure 17: Asymptotic behavior of the overall load in the Heavy-Traffic (top panel), Large-System (middle panel) and Balanced-System (bottom panel) regimes.

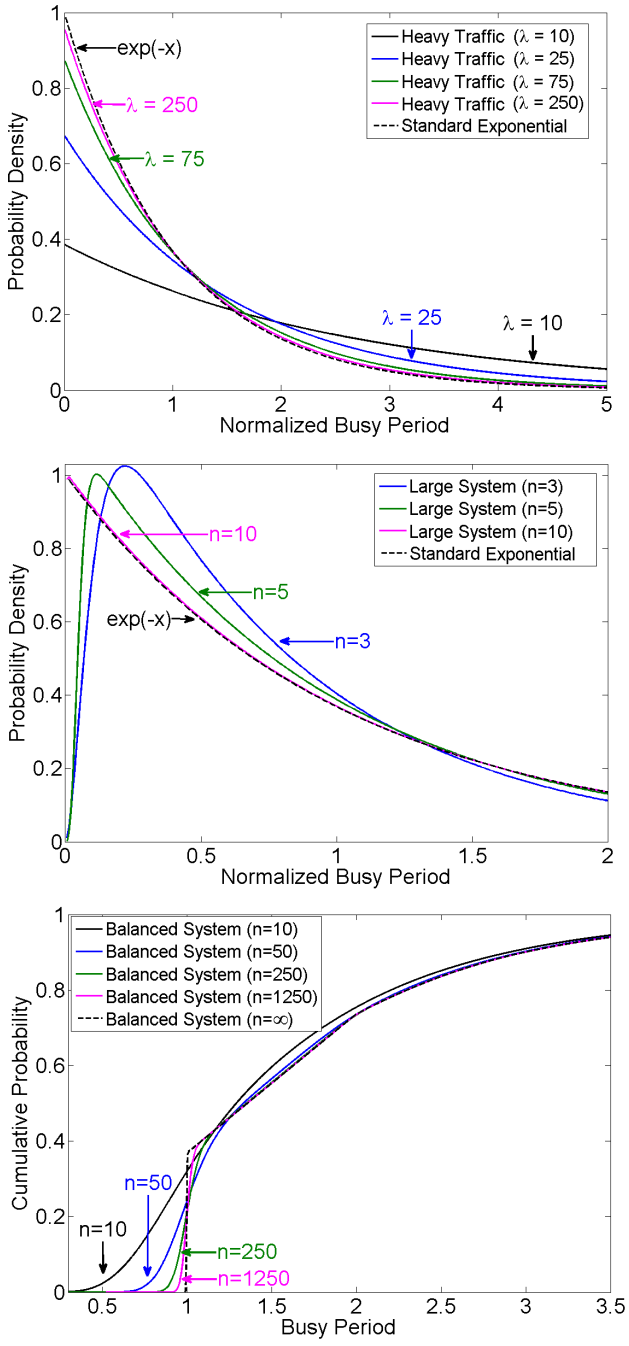

Figure 18: Asymptotic behavior of the busy period the Heavy-Traffic (top panel), Large-System (middle panel), and Balanced-System (bottom panel) regimes; "Standard exponential" is a shorthand for the exponential distribution with unit mean.

In the bottom panel of Figure 19 we plot cumulative distribution functions of the draining time in the Balanced-System regime. Convergence to the asymptotic cumulative distribution function predicted by integrating over the density in Eq. (125) with $\lambda = \tau = 1$, while taking into account the atom at zero (i.e., the probability that $D = 0$) is clearly visible. Similarly, cumulative distribution function of the first occupied site are plotted in the inset and are shown to converge to the asymptotic cumulative distribution function predicted by Eq. (122).

6.4 Asymptotic analysis: The general case

In this section we shift from homogeneous ASIPs to general (inhomogeneous) ASIPs, and extend the stochastic limit laws established in Section 6.2 to the general case. Throughout this section we denote by $m_k = 1/\mu_k$ the mean sojourn time of particles in site k, by $\mathcal{E}$ an exponentially distributed random variable with unit mean, and by Z a Gauss-distributed random variable with zero mean and unit variance.

6.4.1 Heavy traffic

We remind the reader that the heavy-traffic regime considers ASIP lattices in which the inflow rate tends to infinity: $\lambda \to \infty$. Throughout this subsection we set

$$\langle m \rangle = \frac{1}{n} \sum_{k=1}^{n} m_k \tag{128}$$

and

$$\left\langle m^2 \right\rangle = \frac{1}{n} \sum_{k=1}^{n} m_k^2 \ . \tag{129}$$

The ASIP stochastic limit laws — under the heavy-traffic regime — are as follows:

Traversal Time. As is clear from Eq. (91) the inflow rate does not affect the traversal time T. The traversal time is a sum of n independent exponential random variables with corresponding means $\{m_1, \cdots, m_n\}$. Consequently, the Laplace transform of the traversal time is given by

$$\mathbf{E}\left[\exp\left(-\theta T\right)\right] = \prod_{k=1}^{n} \frac{1}{1 + m_k \theta} \tag{130}$$

$(\theta \geq 0)$.

Overall Load. Increasing the inflow rate λ is expected to result in an increase of the overall load L. And indeed, Eq. (93) implies that the mean of the overall load L scales linearly with λ. Consequently, we normalize the overall load L by the dimensionless term $\langle m \rangle \lambda$ and analyze the stochastic limit of the normalized overall load $L/\left(\langle m \rangle \lambda\right)$ (as $\lambda \to \infty$). Setting $z = \exp(-\theta/\left(\langle m \rangle \lambda\right))$ in Eq. (94) we obtain the limit

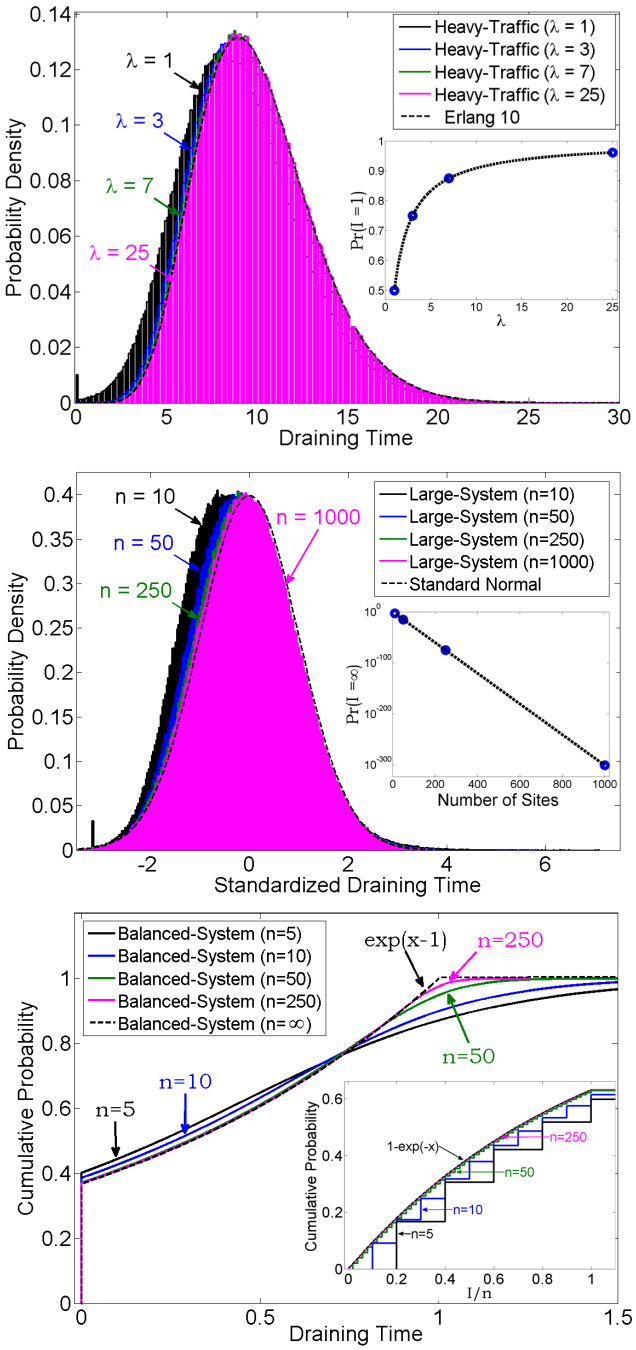

Figure 19: Asymptotic behavior of the first occupied site and draining time in the Heavy-Traffic (top panel), Large-System (middle panel), and Balanced-System (bottom panel) regimes.

$$\lim_{\lambda\to\infty} \mathbf{E}\left[\exp\left(-\theta\frac{L}{\langle m\rangle\,\lambda}\right)\right] = \prod_{k=1}^{n}\frac{1}{1+\frac{m_k}{\langle m\rangle}\theta} \tag{131}$$

($\theta \geq 0$). Eq. (131) implies that the limiting overall load is equal, in law, to the sum of n independent exponential random variables with corresponding means $\{m_1/\langle m\rangle, \cdots, m_n/\langle m\rangle\}$.

Busy Period. As in the case of the overall load, increasing the inflow rate λ is expected to result in an increase of the duration of the busy period B. And indeed, Eq. (98) implies that the mean of the busy period B scales like λ^{n-1}. Consequently, we normalize the busy period B by the dimensionless term $(\langle m\rangle\,\lambda)^{n-1}$ and analyze the stochastic limit of the normalized busy period $B/(\langle m\rangle\,\lambda)^{n-1}$ (as $\lambda\to\infty$). Using Eq. (99) we obtain the limit

$$\lim_{\lambda\to\infty} \mathbf{E}\left[\exp\left(-\theta\frac{B}{(\langle m\rangle\,\lambda)^{n-1}}\right)\right] = \frac{1}{1+\left(\langle m\rangle\prod_{k=1}^{n}\frac{m_k}{\langle m\rangle}\right)\theta} \tag{132}$$

($\theta \geq 0$). Eq. (132) implies that the limiting busy period is equal, in law, to an exponential random variable with mean $\langle m\rangle\prod_{k=1}^{n}(m_k/\langle m\rangle)$. Note that $\prod_{k=1}^{n}\frac{m_k}{\langle m\rangle} \leq 1$ due to the inequality of arithmetic and geometric means.

First Occupied Site. Increasing the inflow rate λ is expected to increase to one the probability of finding the first site occupied. And, indeed, Eq. (102) yields the limit

$$\lim_{\lambda\to\infty} \Pr(I=1) = 1\ . \tag{133}$$

Draining Time. Equation (133) implies that for large λ the first occupied site is effectively the first site. Consequently, for large λ the draining time D should coincide with the traversal time T. And indeed, taking the limit $\lambda\to\infty$ in Eq. (105) confirms this conjecture.

6.4.2 Large Systems

We remind the reader that the large-system regime considers ASIPs in which the number of sites increases to infinity: $n\to\infty$. In Subsection 6.2.2 we analyzed the large-system limit of homogeneous ASIP lattices. Throughout our analysis we have encountered sums of IID random variables and, in turn, applied the classic Central Limit Theorem. In this subsection we will make use of Lyapunov's Central Limit Theorem, a variant of the classical Central Limit Theorem in which the random summands $\{\xi_k\}$ are independent, but not necessarily identically distributed [73]. Lyapunov's theorem requires that there exists some $\delta > 0$ for which the moments of order $(2+\delta)$ of the random variables $\{|\xi_k|\}$ exist and that the rate of growth of these moments is limited by the condition

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n}\mathbf{E}\left[|\xi_k - \mathbf{E}[\xi_k]|^{2+\delta}\right]}{\left(\sum_{k=1}^{n}\mathbf{Var}[\xi_k]\right)^{\frac{2+\delta}{2}}} = 0\,. \tag{134}$$

The theorem then asserts that the sum

$$\sum_{k=1}^{n} \frac{\xi_k - \mathbf{E}\left[\xi_k\right]}{\sqrt{\sum_{k=1}^{n} \mathbf{Var}\left[\xi_k\right]}} \tag{135}$$

converges in distribution to a standard normal random variable Z, as n tends to infinity. A note regarding Lyapunov's condition appears in the Appendix to this chapter.

Throughout this sections we will assume that the random variables $\{\triangle_1, \triangle_2, \triangle_3 \cdots\}$ and $\{G_1, G_2, G_3 \cdots\}$, that were defined, respectively, in subsections 6.1.1 and 6.1.2 obey Lyapunov's condition. In addition, we will assume that

$$\sum_{n=1}^{\infty} \left[\prod_{k=1}^{n-1} (1 + \lambda m_k)^{-1} \sum_{k=1}^{n} m_k \right] < \infty \,. \tag{136}$$

A note regarding the condition in Eq. (136) appears in the Appendix to this chapter. Provided that the conditions in Eqs. (134) and (136) jointly hold, the ASIP stochastic limit laws — under the large-system regime — are as follows:

Traversal Time. Equation (91) asserts that the traversal time T is a sum of n independent exponential random variables, $\{\triangle_1, \cdots, \triangle_n\}$, with the corresponding means $\{m_1, \cdots, m_n\}$ and variances $\left\{m_1^2, \cdots, m_n^2\right\}$. Applying Lyapunov's Central Limit Theorem we obtain that the traversal time T admits the Gaussian stochastic approximation

$$T \approx \sum_{k=1}^{n} m_k + \sqrt{\sum_{k=1}^{n} m_k^2} \cdot Z \tag{137}$$

(as $n \to \infty$).

Overall Load. Equation (94) asserts that the overall load L is a sum of n independent geometric random variables, $\{G_1, \cdots, G_n\}$, with the corresponding means $\{\lambda m_1, \cdots, \lambda m_n\}$ and variances $\left\{\lambda m_1 + (\lambda m_1)^2, \cdots, \lambda m_n + (\lambda m_n)^2\right\}$. Applying Lyapunov's Central Limit Theorem we obtain that the overall load L admits the Gaussian stochastic approximation

$$L \approx \sum_{k=1}^{n} \lambda m_k + \sqrt{\sum_{k=1}^{n} \left(\lambda m_k + (\lambda m_k)^2\right)} \cdot Z \tag{138}$$

(as $n \to \infty$).

Busy Period. Increasing the lattice size n is expected to result in an increase in the length of the busy period. Indeed, Eq. (98) implies that for large n the mean of the busy period scales like $\prod_{k=1}^{n} (1 + \lambda m_k)$. Consequently, we

analyze the stochastic limit of the normalized busy period $B/\prod_{k=1}^{n}(1+\lambda m_k)$ (as $n \to \infty$). Using Eq. (99) we obtain the limit

$$\lim_{n\to\infty} \mathbf{E}\left[\exp\left(-\theta \frac{B}{\prod_{k=1}^{n}(1+\lambda m_k)}\right)\right] = \frac{\lambda}{\lambda+\theta} \tag{139}$$

($\theta \geq 0$). Since the right-hand-side of Eq. (139) is the Laplace transform of an exponential distribution with mean $1/\lambda$, we obtain that the busy period B admits the stochastic approximation

$$B \approx \prod_{k=1}^{n}(1+\lambda m_k) \cdot \frac{1}{\lambda}\mathcal{E} \tag{140}$$

(as $n \to \infty$). The derivation of Eq. (139) is given in the Appendix to this chapter.

First Occupied Site. Taking the limit $n \to \infty$ in Eq. (102) yields

$$\lim_{n\to\infty} \Pr(I = k) = \frac{\lambda m_k}{1+\lambda m_k} \prod_{j=1}^{k-1} \frac{1}{1+\lambda m_j} \tag{141}$$

($k = 1, 2, 3, \cdots$). This result can be interpreted as an inhomogeneous geometric law. The derivation of Eq. (141) is given in the Appendix to this chapter.

Draining Time. In the Appendix to this chapter we show that the regularity condition given in Eq. (136) asserts that $\mathbf{E}\left[\sum_{k=1}^{I-1} \triangle_k\right]$ is a finite constant that does not depend on n. Consequently — since the number of sites tends to infinity ($n \to \infty$) — the draining time D effectively equals the traversal time T. Combining this observation together with Eq. (137) implies that the draining time D admits the Gaussian stochastic approximation

$$D \approx \sum_{k=1}^{n} m_k + \sqrt{\sum_{k=1}^{n} m_k^2} \cdot Z \tag{142}$$

(as $n \to \infty$). The derivation of Eq. (142) is given in the Appendix to this chapter.

6.4.3 Balanced Systems

We remind the reader that the balanced-system regime considers ASIPs in which the number of sites tends to infinity ($n \to \infty$), and the mean sojourn time at each site tends to zero ($m_k \to 0$ for all k). In the case of homogeneous ASIPs the balance between the large number of sites and the rapid gate-opening rates

was attained by setting $m_k = \tau/n$ where τ is an arbitrary positive parameter. In the case of general ASIPs the balance is attained by setting

$$m_k = \phi\left(\frac{k}{n}\right)\frac{1}{n} \tag{143}$$

$(k = 1, \cdots, n)$, where $\phi(u)$ is an arbitrary positive-valued function defined on the unit interval $(0 \leq u \leq 1)$. The integrability conditions that the function $\phi(u)$ needs to meet are $\int_0^1 \phi(u)\, du < \infty$ and $\int_0^1 \phi(u)^2\, du < \infty$. In what follows, and without loss of generality, we further set $\int_0^1 \phi(u)\, du = \tau$.

Applying this balanced-system setting, and taking the limit $n \to \infty$, the following results are obtained: (i) the traversal time T admits the limit of Eq. (119); (ii) the overall load L admits the limit given by Eq. (120); (iii) the busy period B admits the limit given by Eq. (121); (iv) the draining time D admits the limit given by Eq. (124). Namely, in the balanced-system regime, the aforementioned observables — traversal time, overall load, busy period, and draining time — admit the same stochastic limit laws both in the case of homogeneous ASIPs and in the case of general ASIPs. A difference between homogeneous and general ASIPs is displayed by the first occupied site I. Indeed, setting $\hat{I} = I/n$ to be the scaled first occupied site, we obtain the limits

$$\begin{cases} \lim_{n\to\infty} \Pr\left(\hat{I} > x\right) = \exp\left(-\lambda \int_0^x \phi(u)\, du\right) , \\ \\ \lim_{n\to\infty} \Pr(\hat{I} = \infty) = \exp\left(-\lambda\tau\right) , \end{cases} \tag{144}$$

$(0 \leq x \leq 1)$. We note that the above mentioned results can also be obtained under milder assumptions and we refer the reader to the Appendix to this chapter for details and proofs.

As in the homogeneous setting, the general balanced-system limiting regime can be understood as an $M/D/\infty$ queue. Indeed, particles arrive to the lattice following a Poisson process with rate λ. Each particle, upon its arrival to the lattice, starts traversing it. Particles' traversal times are deterministic and of common length τ, and upon "traversal completion" the particles leave the lattice. As in the homogeneous setting, the common deterministic traversal times assure that particles will leave the lattice *exactly* τ units of time after their respective arrival epochs, and will do so in a FIFO manner. One should however note the following difference between the homogeneous and inhomogeneous settings. While in the homogeneous setting particles traverse the lattice at a "constant velocity", in the inhomogeneous setting particles do so with a "local velocity" that depends on their position along the lattice. Specifically, in the homogeneous ASIP the traversal velocity is position-independent and equals $1/\tau$ (in units length per unit time), whereas in the inhomogeneous ASIP the traversal velocity is position-dependent and is given by the function $1/\phi(u)$.

6.5 Conclusions

In this chapter we established stochastic limit laws for five key observables of the ASIP: Traversal Time, Overall Load, Busy Period, First Occupied Site, and Draining Time. We considered three different asymptotic limiting regimes: the heavy-traffic regime, the large-system regime, and the balanced-system regime. We showed that each of these limiting regimes yields a set of stochastic limit laws for the ASIP's five key observables. Each set of limit laws established is, in effect, a characteristic "finger print" of the asymptotic limiting regime applied. The results were obtained analytically and in closed form, and cover both homogeneous and inhomogeneous ASIPs. In the following chapter, we will introduce the reader to Catalan's trapezoids a combinatorial construct instrumental to the analysis of the ASIP.

6.6 Appendix

6.6.1 Proof of the Distributional Little's Law

Let $A(t)$ denote the number of Poisson arrivals during a time interval of length t. Then

$$\mathbf{E}\left[z^{A(T)}\right] = \mathbf{E_T}\left[\mathbf{E}\left[z^{A(T)}|T\right]\right] = \mathbf{E_T}\left[e^{-\lambda(1-z)T}\right], \tag{145}$$

where in the second equality we have used fact that $A(t)$ follows the Poisson distribution with mean λt. The right hand side of Eq. (145) is the Laplace transform of the traversal time T evaluated at the point $\theta = \lambda(1-z)$ and by use Eq. (91) we therefore have

$$\mathbf{E}\left[z^{A(T)}\right] = \prod_{k=1}^{n} \frac{\mu_k}{\mu_k + \lambda(1-z)}. \tag{146}$$

Comparing this result with Eq. (94) it readily follows that $\mathbf{E}\left[z^{L}\right] = \mathbf{E}\left[z^{A(T)}\right]$.

6.6.2 Derivation of Eq. (99)

Considering Eq. (97) and utilizing the law of total expectation we write the Laplace transform of B as

$$\begin{cases} \mathbf{E}\left[\exp\left(-\theta B\right)\right] = Pr(T < \triangle_0)\mathbf{E}\left[\exp\left(-\theta T\right)|T < \triangle_0\right] \\ +Pr(\triangle_0 \leq T)\mathbf{E}\left[\exp\left(-\theta(\triangle_0 + B')\right)|\triangle_0 \leq T\right]. \end{cases} \tag{147}$$

The first term in Eq. (147) is treated by noting that the independence of the random variables $\triangle_0$ and T implies

$$\mathbf{E}\left[\exp\left(-\theta T\right)|T < \triangle_0\right] = \frac{\int_0^\infty f(t)e^{-\theta t}Pr(t < \triangle_0)dt}{Pr(T < \triangle_0)}, \tag{148}$$

where $f(t)$ is the probability density function of T. Since $\triangle_0$ is exponentially distributed with rate λ, $Pr(\triangle_0 > t) = e^{-\lambda t}$, and we have

$$\mathbf{E}\left[\exp\left(-\theta T\right)|T < \triangle_0\right] = \frac{\mathbf{E}\left[\exp\left(-(\theta+\lambda)T\right)\right]}{Pr(T < \triangle_0)}. \tag{149}$$

We now note that the Laplace transform of the random variable T is given by Eq. (91) and we therefore have

$$Pr(T < \triangle_0)\mathbf{E}\left[\exp\left(-\theta T\right)|T < \triangle_0\right] = \prod_{k=1}^{n} \frac{\mu_k}{\mu_k + \theta + \lambda}. \tag{150}$$

The second term in Eq. (147) is treated by noting that the random variables $\{\triangle_0, T, B'\}$ are independent, and that B' is an IID copy of B. Therefore,

$$\begin{aligned}&\mathbf{E}\left[\exp\left(-\theta(\triangle_0+B')\right)|\triangle_0\leq T\right]\\&=\frac{\mathbf{E}[\exp(-\theta B)]\int_0^\infty f(t)\left[\int_0^t g(z)e^{-\theta z}dz\right]dt}{Pr(\triangle_0\leq T)}\,,\end{aligned}\tag{151}$$

where $g(z)=\lambda e^{-\lambda z}$ is the probability density function of $\triangle_0$. The double integral gives:

$$\frac{\lambda}{\lambda+\theta}\int_0^\infty f(t)\left[1-e^{-(\lambda+\theta)t}\right]dt=\frac{\lambda}{\lambda+\theta}\left[1-\prod_{k=1}^{n}\frac{\mu_k}{\mu_k+\theta+\lambda}\right]\,,\tag{152}$$

and we conclude that

$$\begin{aligned}&\mathbf{E}\left[\exp\left(-\theta(\triangle_0+B')\right)|\triangle_0\leq T\right]\\&=\frac{\lambda\mathbf{E}[\exp(-\theta B)]\left[1-\prod_{k=1}^{n}\frac{\mu_k}{\mu_k+\theta+\lambda}\right]}{Pr(\triangle_0\leq T)(\lambda+\theta)}\,.\end{aligned}\tag{153}$$

Substituting Eqs. (150) and (153) into Eq. (147) and rearranging terms we obtain Eq. (99). Equation (98) can be obtained directly by using $\mathbf{E}[B]=-\frac{d\mathbf{E}[\exp(-\theta B)]}{d\theta}|_{\theta=0}$.

6.6.3 Derivation of Eqs. (104) and (105)

Considering Eq. (103) and conditioning on the value of the first non-empty site I we obtain the following expressions:

$$\begin{cases}E[D]=E[E[D|I]]=\sum_{k=1}^{n}Pr(I=k)\sum_{j=k}^{n}\frac{1}{\mu_j}\\E[e^{-\theta D}]=E[E[e^{-\theta D}|I]]=Pr(I=\infty)+\sum_{k=1}^{n}Pr(I=k)\prod_{j=k}^{n}\frac{\mu_j}{\theta+\mu_j}\end{cases}\tag{154}$$

Equations (104) and (105) follow by substituting Eq. (102) into Eq. (154).

6.6.4 Derivation of Eq. (118)

Intuitively, Eq. (118) is most easily understood by noting that in the large-system limit, the draining and traversal times are both given by infinite sums of independent exponential random variables, $D=\sum_{k=I}^{\infty}\triangle_k$ and $T=\sum_{k=1}^{\infty}\triangle_k$, correspondingly. Moreover, the only difference between the two infinite sums is a sum of $I-1$, independent, exponential random variables whose expected value is

$$\lim_{n\to\infty}\mathbf{E}\left[\sum_{k=1}^{I-1}\triangle_k\right]=\mathbf{E_I}\left[\mathbf{E}\left[\sum_{k=1}^{I-1}\triangle_k|I\right]\right]=\frac{1}{\lambda}\,,\tag{155}$$

a finite constant that does not depend on n. The difference between the traversal time and the draining time is hence negligible in the large-system limit.

More precisely, Eq. (118) is derived by substituting $-i\theta$ for θ in (see in the sequel) Eq. (165) to obtain the characteristic function of the draining time

$$\mathbf{E}\left[\exp\left(i\theta D\right)\right] = \frac{-i\theta}{-i\theta-\lambda}\left(\frac{1}{1+\lambda m}\right)^n + \frac{\lambda}{\lambda+i\theta}\left(\frac{1}{1-i\theta m}\right)^n . \tag{156}$$

The characteristic function of the standardized draining time, $\frac{D-nm}{m\sqrt{n}}$, follows

$$\begin{array}{c}\mathbf{E}\left[\exp\left(i\theta\left[\frac{D-nm}{m\sqrt{n}}\right]\right)\right] = \left[\frac{-i\theta/(m\sqrt{n})}{-i\theta/(m\sqrt{n})-\lambda}\left(\frac{1}{1+\lambda m}\right)^n\right. \\ \left. +\frac{\lambda}{\lambda+i\theta/(m\sqrt{n})}\left(\frac{1}{1-i\theta/\sqrt{n}}\right)^n\right]\exp\left(-i\theta\sqrt{n}\right) .\end{array} \tag{157}$$

Recalling the Taylor expansion

$$n\cdot ln\left[\frac{1}{1-i\theta/\sqrt{n}}\right] = i\theta\sqrt{n}-\theta^2/2+O(1/\sqrt{n}) \tag{158}$$

and taking the large-system limit of Eq. (157) we obtain

$$\lim_{n\to\infty}\mathbf{E}\left[\exp\left(i\theta\left[\frac{D-nm}{m\sqrt{n}}\right]\right)\right] = \exp\left(-\theta^2/2\right) , \tag{159}$$

which is the characteristic function of a normal random variable with zero mean and unit variance.

6.6.5 Derivation of Eq. (122)

We now note that

$$\Pr\left(\hat{I}>x\right) = \sum_{i/n>x}\frac{\lambda}{n/\tau+\lambda}\frac{1}{\left(1+\lambda\tau/n\right)^{i-1}} + \Pr\left(\hat{I}=\infty\right) \tag{160}$$

taking the limit $n\to\infty$ we find that the first term is a Riemann sum that converges to the integral

$$\lim_{n\to\infty}\Pr\left(\hat{I}>x\right) = \lambda\tau\int_x^1\exp\left(-\lambda\tau u\right)du . \tag{161}$$

and that the second term is given by

$$\lim_{n\to\infty}\Pr(\hat{I}=\infty) = \lim_{n\to\infty}\left(\frac{1}{1+\lambda\tau/n}\right)^n = \exp\left(-\lambda\tau\right) . \tag{162}$$

Eq. (122) readily follows.

6.6.6 Derivation of Eq. (124)

In order to derive Eq. (124) we first note that, in the case of a homogeneous ASIP lattice, Eq. (105) reads

$$\begin{aligned}\mathbf{E}\left[\exp\left(-\theta D\right)\right] &= \left(\tfrac{1}{1+\lambda m}\right)^n \\ &+ \tfrac{\lambda m}{1+\lambda m}\sum_{k=1}^{n}\left(\tfrac{1}{1+\lambda m}\right)^{k-1}\left(\tfrac{1}{1+\theta m}\right)^{n-k+1}.\end{aligned} \tag{163}$$

We sum the series by noting that

$$\sum_{k=1}^{n} a^{k-1}b^{n-k+1} = \frac{b\left(a^n-b^n\right)}{a-b}, \tag{164}$$

a formula that is easily proved by use of either geometric series summation or mathematical induction. We obtain

$$\mathbf{E}\left[\exp\left(-\theta D\right)\right] = \frac{\theta}{\theta-\lambda}\left(\frac{1}{1+\lambda m}\right)^n + \frac{\lambda}{\lambda-\theta}\left(\frac{1}{1+\theta m}\right)^n. \tag{165}$$

Equation (124) follows from substituting $m = \tau/n$ into Eq. (165) and taking the limit $n \to \infty$.

6.6.7 Derivation of the Large System Limiting Regime — General Case

- ***Notes on Regularity Conditions***

1. In practice it is usually easiest to check the Lyapunov's condition for $\delta = 1$ and it is easily verified that the condition holds for the *special case* in which the following two limits exist:

$$\begin{cases} \left\langle \sigma^2 \right\rangle = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{Var}\left[\xi_k\right], \\ \left\langle \kappa^3 \right\rangle = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}\left[\left|\xi_k - \mathbf{E}\left[\xi_k\right]\right|^3\right]. \end{cases} \tag{166}$$

2. The fact that Lyapunov's condition holds for the random variables $\{\triangle_1, \triangle_2, \triangle_3 \cdots\}$ and $\{G_1, G_2, G_3 \cdots\}$ implies that

$$\begin{cases} \sum_{k=1}^{\infty} m_k = \infty, \\ \prod_{k=1}^{\infty} \left(1+\lambda m_k\right) = \infty. \end{cases} \tag{167}$$

Indeed, since $\sum_{k=1}^{n} \mathbf{E}\left[\left|\xi_k - \mathbf{E}\left[\xi_k\right]\right|^{2+\delta}\right]$ is monotonically increasing with n, Eq. (134) implies that $\sum_{k=1}^{\infty} \mathbf{Var}\left[\xi_k\right]$ diverges. In the case of the random variables, $\{\triangle_1, \triangle_2, \triangle_3 \cdots\}$, this means that $\sum_{k=1}^{\infty} m_k^2$ diverges and in the case of the random variables, $\{G_1, G_2, G_3 \cdots\}$, this means that $\sum_{k=1}^{\infty} \left(\lambda m_k + \lambda^2 m_k^2\right)$ diverges. In any case, since $\sum_{k=1}^{\infty} m_k < \infty \Rightarrow \sum_{k=1}^{\infty} m_k^2 < \infty$ it follows that $\sum_{k=1}^{\infty} m_k$ must diverge as well. In addition, since

$$1 + \lambda \sum_{k=1}^{n} m_k \leq \prod_{k=1}^{n} \left(1 + \lambda m_k\right) \leq e^{\lambda \sum_{k=1}^{n} m_k}, \tag{168}$$

it follows that $\sum_{k=1}^{\infty} m_k$ and $\prod_{k=1}^{\infty} \left(1 + \lambda m_k\right)$ converge or diverge together.

3. The regularity condition in Eq. (136) implies that

$$\lim_{n \to \infty} \left[\frac{\sum_{k=1}^{n} m_k}{\prod_{k=1}^{n-1} \left(1 + \lambda m_k\right)} \right] = 0 . \tag{169}$$

Since

$$\frac{\frac{m_j}{(1+\lambda m_j)}}{\prod_{k=1}^{n} \left(1 + \lambda m_k\right)} \leq \frac{\sum_{k=1}^{n} \frac{m_k}{(1+\lambda m_k)}}{\prod_{k=1}^{n} \left(1 + \lambda m_k\right)} \leq \frac{\sum_{k=1}^{n} m_k}{\prod_{k=1}^{n-1} \left(1 + \lambda m_k\right)} \tag{170}$$

it follows that

$$\begin{cases} \lim_{n \to \infty} \left[\frac{\frac{m_j}{(1+\lambda m_j)}}{\prod_{k=1}^{n} (1+\lambda m_k)} \right] = 0 , \\ \lim_{n \to \infty} \left[\frac{\sum_{k=1}^{n} \frac{m_k}{(1+\lambda m_k)}}{\prod_{k=1}^{n} (1+\lambda m_k)} \right] = 0 . \end{cases} \tag{171}$$

- ***Busy Period***

From Eq. (99), the Laplace transform $B/\prod_{k=1}^{n}(1+\lambda m_k)$ is given by

$$\mathbf{E}\left[\exp\left(-\theta\frac{B}{\prod_{k=1}^{n}(1+\lambda m_k)}\right)\right] = \frac{\lambda+\theta/\prod_{k=1}^{n}(1+\lambda m_k)}{\lambda+\theta\prod_{k=1}^{n}\left[1+\theta m_k/\left((1+\lambda m_k)\prod_{j=1}^{n}(1+\lambda m_j)\right)\right]}. \tag{172}$$

Equation (167) asserts that the second term in the nominator of the right hand side of Eq. (172) is negligible in the large n limit. Taking the logarithm of the second term in the right hand side of the denominator of Eq. (172) and using Eq. (171) we have

$$log\left[\prod_{k=1}^{n}\left[1+\theta m_k/\left((1+\lambda m_k)\prod_{j=1}^{n}(1+\lambda m_j)\right)\right]\right] \cong \theta\left[\frac{\sum_{k=1}^{n}\frac{m_k}{(1+\lambda m_k)}}{\prod_{k=1}^{n}(1+\lambda m_k)}\right] \longrightarrow 0 \tag{173}$$

(as $n\to\infty$). The result in Eq. (139) follows from the continuity of the exponential function.

- ***First Occupied Site***

In order to obtain Eq. (141) it is enough to take the limit $n\to\infty$ in Eq. (102) and use Eq. (167).

- ***Draining Time***

Provided that Lyapunov's condition holds for the random variables, $\{\triangle_1,\triangle_2,\triangle_3\cdots\}$, Eq. (137) asserts that

$$\frac{\sum_{k=1}^{n}\triangle_k-\sum_{k=1}^{n}m_k}{\sqrt{\sum_{k=1}^{n}m_k^2}} \xrightarrow{d} Z \tag{174}$$

(as $n\to\infty$). In order to show that

$$\frac{D-\sum_{k=1}^{n}m_k}{\sqrt{\sum_{k=1}^{n}m_k^2}} \xrightarrow{d} Z \tag{175}$$

(as $n \to \infty$), we note that $D = \sum_{k=I}^{n} \triangle_k$ and recall that if ξ_n is a random variable that converges in distribution to ξ and the difference between the random variables ξ_n and ζ_n converges in probability to zero, then ζ_n also converges in distribution to ξ [74]. It is therefore sufficient to show that

$$\lim_{n\to\infty} Pr\left(\left|\frac{\sum_{k=1}^{I-1} \triangle_k}{\sqrt{\sum_{k=1}^{n} m_k^2}} > \varepsilon\right|\right) \to 0\ . \tag{176}$$

By use of Markov's inequality we have

$$Pr\left(\left|\frac{\sum_{k=1}^{I-1} \triangle_k}{\sqrt{\sum_{k=1}^{n} m_k^2}} > \varepsilon\right|\right) \leq \frac{\mathbf{E}\left[\sum_{k=1}^{I-1} \triangle_k\right]}{\varepsilon\sqrt{\sum_{k=1}^{n} m_k^2}}\ . \tag{177}$$

By use of Eq. (141) the nominator in the right hand side of Eq. (177) gives

$$\begin{aligned} &\mathbf{E}\left[\sum_{k=1}^{I-1} \triangle_k\right] = \mathbf{E_I}\left[\mathbf{E}\left[\sum_{k=1}^{I-1} \triangle_k | I\right]\right] \cong \\ &\sum_{n=1}^{\infty} \tfrac{\lambda m_n}{1+\lambda m_n} \prod_{k=1}^{n-1} (1+\lambda m_j)^{-1} \sum_{k=1}^{n-1} m_k\ . \end{aligned} \tag{178}$$

Since

$$\frac{\frac{\lambda m_n}{1+\lambda m_n} \sum_{k=1}^{n-1} m_k}{\prod_{k=1}^{n-1} (1+\lambda m_k)} \leq \frac{\sum_{k=1}^{n} m_k}{\prod_{k=1}^{n-1} (1+\lambda m_k)}\ , \tag{179}$$

the regularity condition in Eq. (136) asserts that $\mathbf{E}\left[\sum_{k=1}^{I-1} \triangle_k\right]$ is a finite constant that does not depend on n. Recalling that Lyapunov's condition for the random variables, $\{\triangle_1, \triangle_2, \triangle_3 \cdots\}$ implies that $\sum_{k=1}^{\infty} m_k^2 = \infty$ (see notes on regularity conditions above), we conclude that

$$\lim_{n\to\infty} \frac{\mathbf{E}\left[\sum_{k=1}^{I-1} \triangle_k\right]}{\varepsilon\sqrt{\sum_{k=1}^{n} m_k^2}} = 0 \tag{180}$$

and the desired result, Eq. (142), follows.

6.6.8 Derivation of the Balanced System Limiting Regime — General Case

- ***Regularity Conditions***

In proving the results presented in Section 6.4.3 we first note that the setup depicted there is a special case of a more general setup. Here we will assume that the set $\{m_k(n)\}$ is chosen such that there exists a positive-valued function $\phi(u)$ that obeys

$$\begin{aligned} &\lim_{n\to\infty} \sum_{k/n<x} m_k(n) = \int_0^x \phi(u)du < \infty \\ &\lim_{n\to\infty} \sum_{k/n<x} m_k(n)^2 = 0 \end{aligned} \tag{181}$$

$(0 \leq x \leq 1)$. In particular, and without loss of generality, we denote

$$F(x) = \int_0^x \phi(u)du \tag{182}$$

$(0 \leq x \leq 1)$, and set $F(1) = \tau$. One can now easily verify that Eq. (181) holds for the special case in which $m_k(n) = \phi\left(\frac{k}{n}\right)\frac{1}{n}$, $\int_0^1 \phi(u)\,du = \tau$ and $\int_0^1 \phi(u)^2\,du < \infty$.

- ***Traversal Time***

Taking the logarithm of Eq. (91) we obtain

$$log\left[\mathbf{E}\left[\exp\left(-\theta T\right)\right]\right] = -\sum_{k=1}^{n} log\left[1 + m_k(n)\theta\right] \tag{183}$$

$(\theta \geq 0)$. We now note that

$$-\sum_{k=1}^{n} log\left[1 + m_k(n)\theta\right] \cong -\theta\sum_{k=1}^{n} m_k(n) - \frac{\theta}{2}\sum_{k=1}^{n} m_k(n)^2 \tag{184}$$

and after taking the balanced-system limit of this equation we have

$$\lim_{n\to\infty} -\sum_{k=1}^{n} log\left[1 + m_k(n)\theta\right] = -\theta\tau\,. \tag{185}$$

The desired result, Eq. (119), follows from the continuity of the exponential function.

- ***Overall Load***

Taking the logarithm of Eq. (94) we obtain

$$log\left[\mathbf{E}\left[z^{L}\right]\right] = -\sum_{k=1}^{n} log\left[1 + m_k(n)\lambda\left(1-z\right)\right] \tag{186}$$

($|z| \leq 1$). We now note that

$$\begin{aligned} & -\sum_{k=1}^{n} log\left[1 + m_k(n)\lambda\left(1-z\right)\right] \\ & \cong -\lambda\left(1-z\right)\sum_{k=1}^{n} m_k(n) - \tfrac{\lambda(1-z)}{2}\sum_{k=1}^{n} m_k(n)^2\,, \end{aligned} \tag{187}$$

and after taking the balanced-system limit of this equation we have

$$\lim_{n\to\infty} -\sum_{k=1}^{n} log\left[1 + m_k(n)\lambda\left(1-z\right)\right] = -\lambda\left(1-z\right)\tau\,. \tag{188}$$

The desired result, Eq. (120), follows from the continuity of the exponential function.

- ***Busy Period***

Taking the balanced-system limit of Eq. (99) we have (by use of continuity)

$$\begin{aligned} & \lim_{n\to\infty} \mathbf{E}\left[\exp\left(-\theta B\right)\right] = \\ & \frac{\lambda+\theta}{\lambda+\theta\lim_{n\to\infty}\prod_{k=1}^{n}\left[1+(\lambda+\theta)m_k(n)\right]}\,. \end{aligned} \tag{189}$$

($\theta \geq 0$). We now note that

$$\begin{aligned} & log\left[\prod_{k=1}^{n}\left[1+(\lambda+\theta)m_k(n)\right]\right] = \sum_{k=1}^{n} log\left[1+(\lambda+\theta)m_k(n)\right] \\ & \simeq (\lambda+\theta)\sum_{k=1}^{n} m_k(n) + \tfrac{\lambda+\theta}{2}\sum_{k=1}^{n} m_k(n)^2\,, \end{aligned} \tag{190}$$

and after taking the balanced-system limit of this equation we have

$$\lim_{n\to\infty} log\left[\prod_{k=1}^{n}\left[1+(\lambda+\theta)m_k(n)\right]\right] = (\lambda+\theta)\tau\,. \tag{191}$$

The desired result, Eq. (121), follows from the continuity of the exponential function.

- ***First Occupied Site***

We note that

$$\begin{aligned}\Pr\left(\hat{I} > x\right) &= \Pr\left(\hat{I} = \infty\right) \\ &+ \sum_{k/n>x} \frac{\lambda}{1/m_k(n)+\lambda} \frac{1}{\prod_{j=1}^{k-1}(1+\lambda m_j(n))} \,.\end{aligned} \tag{192}$$

Taking the balanced-system limit of both sides we find that the first term is given by

$$\lim_{n\to\infty} \Pr(\hat{I} = \infty) = \lim_{n\to\infty} \prod_{k=1}^{n} \frac{1}{1+\lambda m_k(n)} = \exp\left(-\lambda\tau\right) \,. \tag{193}$$

The second term converges to the integral

$$\Pr\left(x < \hat{I} \leq 1\right) = \lambda \int_x^1 \phi(u) \exp\left(-\lambda F(u)\right) du \tag{194}$$

which in turn gives

$$\Pr\left(x < \hat{I} \leq 1\right) = \exp\left(-\lambda F(x)\right) - \exp\left(-\lambda\tau\right) \,. \tag{195}$$

Equation (144) readily follows.

- ***Draining Time***

We first note that

$$\mathbf{E}\left[\exp\left(-\theta D\right)\right] = \mathbf{E}_{\hat{\mathbf{I}}}\left[\mathbf{E}\left[\exp\left(-\theta D\right)|\hat{I}\right]\right] \,. \tag{196}$$

Taking the balanced-system limit of both sides we have

$$\begin{aligned}&\lim_{n\to\infty} \mathbf{E}\left[\exp\left(-\theta D\right)\right] = \exp\left(-\lambda\tau\right) \\ &+\lambda \int_0^1 \phi(u) \exp\left(-\lambda F(u)\right) \exp\left(-\theta\left(\tau - F(u)\right)\right) du \,,\end{aligned} \tag{197}$$

and the desired result, Eq. (124), follows.

7 Catalan's Trapezoids

Named after the French-Belgian mathematician Eugène Charles Catalan, Catalan's numbers arise in various combinatorial problems [75]. Catalan's triangle, a triangular array of numbers somewhat similar to Pascal's triangle, extends the combinatorial meaning of Catalan's numbers and generalizes them [76, 77, 78, 79]. A need for a generalization of Catalan's triangle itself arose while conducting a probabilistic analysis of the ASIP. In this chapter, we introduce the reader to *Catalan's trapezoids*, a countable set of trapezoids whose first element is Catalan's triangle. An iterative scheme for the construction of these trapezoids is presented, a closed-form formula for their entries is derived and their combinatorial meaning is further interpreted. Catalan's trapezoids will come to our aid in Chapter 8 where we derive explicit solutions for occupation probabilities in the ASIP model.

7.1 Catalan's Numbers and Catalan's Triangle

Consider a string of numbers composed of n $(+1)$'s and n (-1)'s, arranged in a row from left to right, such that the sum over every initial substring is non-negative. What is the total number of different such strings? Consider equivalently a path that: (i) starts at the origin of a two dimensional lattice; (ii) consists of n right $(\rightarrow)$ steps and n up $(\uparrow)$ steps; (iii) does not go above the line $y = x$. What is the total number of different such paths? As it turns out, the solution to these combinatorial problems is given by the n^{th} Catalan number [75]:

$$C(n) = \left(\begin{array}{c} 2n \\ n \end{array} \right) - \left(\begin{array}{c} 2n \\ n-1 \end{array} \right) \tag{198}$$

$(n = 1, 2, 3, \cdots)$, with $C(0) = 1$ by definition. Specifically, the first Catalan numbers are given by: $1, 1, 2, 5, 14, 42, 132, 429$.

One can generalize the combinatorial problem mentioned above by considering strings of n $(+1)$'s and k (-1)'s or, alternatively, paths of n right steps and k up steps. In this case, the number of different strings for which the sum over every initial substring is non-negative is given by:

$$C(n,k) = \begin{cases} 1 & k = 0 \\ \left(\begin{array}{c} n+k \\ k \end{array} \right) - \left(\begin{array}{c} n+k \\ k-1 \end{array} \right) & 1 \leq k \leq n \\ 0 & k > n \end{cases} \tag{199}$$

$(n = 0, 1, 2, \cdots ; k = 0, 1, 2, \cdots)$, and the same is true for the number of paths that start at the origin of a two dimensional lattice and do not go above the line $y = x$.

n/k	0	1	2	3	4	5	6	7
0	**1**							
1	**1**	1						
2	**1**	2	2					
3	**1**	3	5	5				
4	**1**	4	9	14	14			
5	**1**	5	14	28	42	42		
6	**1**	6	20	48	90	132	132	
7	**1**	7	27	75	165	297	429	429

Table 3: Some entries of Catalan's triangle.

The numbers $C(n,k)$ are referred to in combinatorial mathematics as the entries of *Catalan's triangle* [75, 76, 77, 78]. These entries facilitate the solution to Bertrand's famous ballot problem [59]: "In an election where candidate A receives n votes and candidate B receives k votes, what is the probability that A will not trail behind B throughout the entire count of votes?". Indeed, the answer to this version of Bertrand's problem is given by the ratio $C(n,k)/\left(\begin{array}{c} n+k \\ k \end{array}\right)$.

Catalan's triangle, illustrated in Table 3, has the following iterative construction. By definition, all entries that are positioned on the left boundary of the triangle ($k=0$) are given by the boundary condition $C(n,0)=1$. In Table 3, these entries are highlighted in bold. Entries positioned to the right of the main diagonal $k=n$ are zero. In Table 3, these entries are indicated by empty squares. All the other entries of Catalan's triangle follow the recursion

$$C(n,k) = C(n-1,k) + C(n,k-1)\,, \tag{200}$$

i.e., each entry is a sum of the entry above it and the entry to its left. In Table 3, a specific example, $9+5=14$, is highlighted in magenta. Entries on the diagonal of Catalan's triangle ($k=n$) are the Catalan numbers: $C(n,n)=C(n)$. In Table 3, these entries are highlighted in blue.

The combinatorial meaning of Eq. (200) and its validity for $1 \leq k \leq n$ become immediately clear after conducting a binary partition of all valid strings according to their last digit $+1$ or -1. Indeed, since $k \leq n$ the sum over a string of n $(+1)$'s and k (-1)'s is non-negative. Moreover, if the string ends with $+1$ there are exactly $C(n-1,k)$ ways to choose the order of the first $n-1$ $(+1)$'s and k (-1)'s such that the sum over every initial substring is non-negative. Similarly, if the string ends with a -1 there are exactly $C(n,k-1)$ ways to choose the order of the first n $(+1)$'s and $k-1$ (-1)'s such that the sum over every initial substring is non-negative. Equation (200) readily follows.

7.2 Catalan's Trapezoids

The need for a generalization of Catalan's triangle naturally arose while conducting a probabilistic analysis of the ASIP. Analyzing the ASIP, it so turned out that steady state occupation probabilities in the model can be written in terms of entries taken from trapezoid number arrays whose iterative construction is identical to that of Catalan's triangle, albeit a small change in boundary conditions. Hence, we set out to construct a family of *Catalan trapezoids.*

Let $C_m(n,k)$ denote the (n,k) entry of the Catalan's trapezoid of order m $(m = 1, 2, 3, \cdots)$. Defining Catalan's trapezoid of order $m = 1$ to be Catalan's triangle we have $C_1(n,k) = C(n,k)$. The iterative construction of higher order trapezoids is similar to that of Catalan's triangle. All elements on the left boundary $(k = 0)$ of the trapezoid are given by the boundary condition $C_m(n,0) = 1$, all elements on the upper boundary of the trapezoid $(n = 0; 0 \leq k \leq m-1)$ are given by the boundary condition $C_m(0,k) = 1$, and all elements positioned to the right of the diagonal $k = n+m-1$ are set to zero. The rest of the elements in the trapezoid follow a recursive rule similar to the one given in Eq. (200), albeit replacing the numbers $C(n,k)$ by the numbers $C_m(n,k)$:

$$C_m(n,k) = C_m(n-1,k) + C_m(n,k-1)\,, \tag{201}$$

i.e., each entry is a sum of the entry above it and the entry to its left. Some entries of Catalan's trapezoid of order $m = 2$ and of order $m = 3$ are given in Table 4.

A closed form expression for $C_m(n,k)$ is given by

$$C_m(n,k) = \begin{cases} \left(\begin{array}{c} n+k \\ k \end{array} \right) & 0 \leq k < m \\ \left(\begin{array}{c} n+k \\ k \end{array} \right) - \left(\begin{array}{c} n+k \\ k-m \end{array} \right) & m \leq k \leq n+m-1 \\ 0 & k > n+m-1 \end{cases} \tag{202}$$

$(n = 0, 1, 2, \cdots; k = 0, 1, 2, \cdots; m = 1, 2, 3, \cdots)$. Indeed, substituting Eq. (202) into Eq. (201) and making use of the well known Pascal's rule [59] one can easily verify that the recursion rule in Eq. (201) holds. The validity of the trapezoid boundary conditions can be easily verified as well.

We will now show that $C_m(n,k)$ is the number of different strings of n $(+1)$'s and k (-1)'s for which the sum over every initial substring is larger than, or equal to, a threshold level $1-m$ $(m = 1, 2, 3, \cdots)$. Setting $m = 1$ we note that this combinatorial interpretation generalizes the combinatorial interpretation given for the entries of Catalan's triangle.

In order to prove that our combinatorial interpretation is valid we will consider an equivalent path counting problem. In the non-negative quadrant of a two dimensional lattice $\{(x,y)|x,y = 0, 1, 2, 3, \cdots\}$, what is the total number

n/k	0	1	2	3	4	5	6	7	8
0	**1**	**1**							
1	**1**	2	2						
2	**1**	3	5	5					
3	**1**	4	9	14	14				
4	**1**	5	14	28	42	42			
5	**1**	6	20	48	90	132	132		
6	**1**	7	27	75	165	297	429	429	
7	**1**	8	35	110	275	572	1001	1430	1430

n/k	0	1	2	3	4	5	6	7	8	9
0	**1**	**1**	**1**							
1	**1**	2	3	3						
2	**1**	3	6	9	9					
3	**1**	4	10	19	28	28				
4	**1**	5	15	34	62	90	90			
5	**1**	6	21	55	117	207	297	297		
6	**1**	7	28	83	200	407	704	1001	1001	
7	**1**	8	36	119	319	726	1430	2431	3432	3432

Table 4: Some entries of Catalan's trapezoid of order $m = 2$ (top) and $m = 3$ (bottom). Entries on the left and upper boundaries are marked in bold. Null entries positioned to the right of the diagonal $k = n + m - 1$ are indicated by empty squares. All other entries follow the recursive rule given in Eq. (201). Two specific examples, $429 + 572 = 1001$ and $117 + 83 = 200$, are highlighted in magenta.

of paths that: (i) start at the origin $(0, 0)$; (ii) are composed out of n right steps $(\rightarrow)$ and k up steps $(\uparrow)$; (iii) do not go above the line $y = x + m - 1$ $(m = 1, 2, 3, \cdots)$? The formulation of this path counting problem asserts that if a path meets the above-mentioned requirements then at any point along the path the number of right steps minus the number of up steps is larger than or equal to $1 - m$. Noting the one to one correspondence between $(+1)$'s and right steps and (-1)'s and up steps, it is clear that the path counting problem we have introduced is equivalent to the string counting problem used to combinatorialy interpret the entries of Catalan's trapezoid of order m. Our proof will be concluded by showing that Eq. (202) is the answer to the path counting problem presented above.

Firstly, consider the case $0 \leq k < m$. In this case paths cannot go above the line $y = x + m - 1$, so every path is legitimate and the total number of paths is $\begin{pmatrix} n+k \\ k \end{pmatrix}$. Secondly, consider the case $k > n + m - 1$. In this case all paths will end at a point which is positioned above the line $y = x + m - 1$, thus yielding no legitimate paths. Thirdly, note that when $m \leq k \leq n + m - 1$

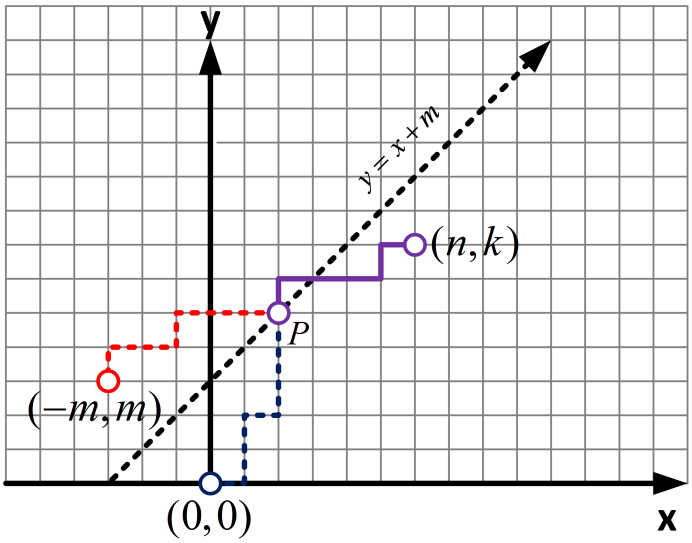

Figure 20: An illustration of the reflection principle ($m = 3$).

some paths will go above the line $y = x + m - 1$ (illegitimate paths) while others will not (legitimate paths). Clearly, the number of legitimate paths is given by the total number of paths minus the number of illegitimate paths. In order to count the number of illegitimate paths we apply a reflection principle.

An illegitimate path connecting the origin with the point (n, k) is illustrated in Figure 20. Every illegitimate path must hit the line $y = x + m$ at least once and we denote the first (leftmost) hitting point by P. The point P divides the illegitimate path into two segments. The path segment positioned to the left of P connects it with the origin (the dashed blue segment in Figure 20). The path segment positioned to the right of P connects it with the point (n, k) (the solid magenta segment in Figure 20). Reflecting the blue segment with respect to a mirror plane placed along the line $y = x + m$ results in a new path segment that connects the point $(-m, m)$ with the point P (the dashed red segment in Figure 20). Concatenating the red segment with the magenta segment results in a semi-reflected path that connects the point $(-m, m)$ with the point (n, k) via P. Since $k \leq n + m - 1$, the point (n, k) lies below the line $y = x + m$ and hence every path that starts at $(-m, m)$ and ends at (n, k) must cross this line at least once. Denoting the first (leftmost) crossing point by P asserts a one to one correspondence between illegitimate paths and paths that: (i) start at $(-m, m)$; (ii) are composed of $n + m$ right steps and $k - m$ up steps. The number of illegitimate paths is thus given by $\begin{pmatrix} n+k \\ k-m \end{pmatrix}$. In turn, since the total number of paths is $\begin{pmatrix} n+k \\ k \end{pmatrix}$, we conclude that the number of legitimate

paths is given by $\begin{pmatrix} n+k \\ k \end{pmatrix} - \begin{pmatrix} n+k \\ k-m \end{pmatrix}$.

7.3 A Generalized Ballot Problem

Consider a generalized ballot problem in which candidate A begins the race $m-1$ votes ahead of candidate B ($m = 1, 2, 3, \cdots$), and collects n more votes for a total of $n+m-1$ to B's k votes ($n = 0, 1, 2, \cdots ; k = 0, 1, 2, \cdots$). What is the probability that candidate A will not trail behind candidate B throughout the entire count of votes? We note in passing that an equivalent problem is one in which the voting starts off with no head start, and the probability that candidate A will not trail behind candidate B, by more than $m-1$ votes, is the one of interest.

Catalan's trapezoids facilitate a solution to the above-mentioned generalization of the ballot problem. Indeed, this can be seen by shifting Figure 20's path and reflection line $m-1$ units to the right, for a path from $(m-1, 0)$ to $(n+m-1, k)$ and a semi-reflected path starting at $(-1, m)$. The reflecting boundary line is now given by $y = x+1$ and its crossing deems a path illegitimate from the stance of B's tally exceeding A's at the point of reflection (even with A's initial vote lead). It is thus clear that the solution to the problem is precisely $C_m(n,k) / \begin{pmatrix} n+k \\ k \end{pmatrix}$.

7.4 Conclusions

In this chapter, we have introduced Catalan's trapezoids — a countable set of number arrays which generalize Catalan's numbers and Catalan's triangle. Catalan's trapezoids facilitate the solution to a generalized ballot problem but the real motivation behind their development comes from the fact that they are an essential tool in the combinatorial analysis of the ASIP. In the following chapter, we will show that steady state occupation probabilities in the model can be written in terms of entries taken from Catalan's trapezoids. The emergence of these combinatorial quantities is of course not coincidental and, as we will hereby show, can be traced back to a certain path counting problem that lurks behind the scenes. To this end, it is extremely interesting to note that Catalan's numbers — and the entries of Catalan's triangle (a.k.a ballot numbers) — are also known to appear in the exact solution for the steady state probability distribution of the ASEP where similar path counting problems arise [23, 80]. We are currently unaware of any previous appearances of Catalan's trapezoids but hope that in similarity to Catalan's numbers they too will find numerous applications.

8 Occupation Probabilities and Fluctuations

Even the simplest ASIPs — homogeneous ASIPs, in which gate opening rates do not depend on the position along the lattice — were shown to display an intriguing showcase of complexity, including power law occupations statistics, diverse forms of self-similarity, and a rich limiting behavior (see Chapter 4). However, several of the aforementioned 'complexity results' relied only on Monte-Carlo studies, as an exact expression for the joint stationary probability distribution of particle occupations is not known (see Chapter 5). Moreover, obtaining such an expression is undoubtedly difficult as coalescence introduces strong correlations between the occupations of different lattice sites.

The main goal of this chapter is to present an exact closed-form expression for the probability that a given number of particles occupies a given set of consecutive lattice sites on an homogeneous ASIP lattice. These probabilities, which we term the incremental load probabilities (to be defined precisely below), are marginals of the joint occupation distribution. Progress can be made with their analysis by using the empty-interval method, a method which has proven useful in the study of aggregation in closed systems [54, 55]. The actual calculation of the probabilities in our open system is based on a combinatorial analysis of the incremental load and on the solution of a boundary value problem that governs its distribution. This approach yields exact, closed form, results expressed in terms of the entries of Catalan's trapezoids — number arrays which were introduced in Chapter 7.

The incremental load probabilities provide valuable information on the ASIP steady state, and furnish an analytical proof for several of the numerical results obtained in Chapter 4. In particular, we prove that: (i) the probability that the k^{th} lattice site is non-empty decays like $1/\sqrt{k}$; (ii) the variance of the occupancy of the k^{th} lattice site grows like $\sqrt{k}$; and (iii) the ASIP's outflow is governed by Rayleigh-distributed inter-exit times.

Before presenting the exact expression for the incremental load probabilities, we follow a complementary approach which is based on mapping the original problem onto its diffusion limit counterpart. This approach shows that the incremental load probabilities in lattice segments that are far away downstream have asymptotic scaling forms which we compute. Some of these scaling forms were previously found in Ref. [56] using an alternative discrete approach. Here we present a "real-space" analysis performed in a continuum limit. Our analysis yields physical insight into the behavior of the model and allows us to derive some new asymptotic scaling forms. More importantly, the diffusion-limit approach reveals that the asymptotics of the incremental load probabilities are *universal*, in the sense that they do not depend on the details of the process which feeds particles into the ASIP lattice.

The chapter is organized as follows. The main results of this chapter are summarized in Section 8.1 where we also introduce the notion of incremental load. In Section 8.2 the ASIP is described as a coagulation model and the empty-interval method is adapted to its analysis. In Section 8.3 a continuum diffusion limit is carried out and asymptotic results are obtained; various implications

of these results are discussed in Section 8.4. Section 8.5 further deepens the probabilistic analysis of the incremental load, and the associated boundary-value problem. In this section we obtain expressions for the incremental load which may be efficiently computed even for inhomogeneous ASIPs. In Section 8.6 we return to homogeneous systems, for which we solve the boundary value problem and obtain a set of exact, closed form, results. Section 8.7 concludes the chapter with an overview and future outlook.

8.1 A summary of key results

In this section we present a short summary of the key results to be established in this chapter. Some of the results presented herein were previously observed in numerical simulations (see Chapter 4). Here, we derive these results analytically and considerably generalize them. In what follows we consider a homogeneous ASIP with $\mu_1 = \cdots = \mu_n = \mu$, and set X_k to be a random variable which represents the fluctuating number of particles present in site k in the steady state. We open this section with a series of asymptotic (large k) results for the distribution and moments of X_k. The asymptotic results below all stem from the main result of this chapter — an exact derivation of the steady-state distribution of the ASIP's incremental load — with which we conclude this section.

Occupation probabilities. The Monte Carlo simulations presented in Chapter 4 concluded that the probability that site k is occupied, $\Pr(X_k > 0)$, decays like $1/\sqrt{k}$ (as $k \to \infty$). Here, we analytically prove that

$$\Pr(X_k > 0) = 1 - \Pr(X_k = 0) \simeq \frac{1}{\sqrt{\pi k}} \ , \tag{203}$$

where "$\simeq$" denotes asymptotic equivalence to leading order in k. We further obtain a scaling form for the probability that site k is occupied by $1 \ll l \ll k$ particles:

$$\Pr(X_k = l) \simeq \frac{\mu}{\lambda k} \phi\left(\frac{\mu l}{\lambda \sqrt{k}}\right) \ , \tag{204}$$

where

$$\phi(u) = \frac{1}{\sqrt{4\pi}} u e^{-u^2/4}. \tag{205}$$

We note that the result of Eq. (203) — contrary to the result of Eq. (204) — is universal in the sense that it does not depend on the arrival rate λ. In fact, we show that this result is universal in a wider sense, and that Eq. (203) holds for any particle arrival process (not necessarily Poissonian). A similar, although slightly weaker, universality holds for the result in Eqs. (204)–(205): while the scaling variable u depends on the arrival rate, the scaling function (205) does not depend on the details of the arrival process. The extent to which this claim is correct is discussed, along with other universality related issues, in Section 8.3.

Conditional mean occupancy. In Chapter 5 it was shown that in homogeneous ASIPs the mean occupancy of site k at steady state is given by

$$\langle X_k \rangle = \lambda/\mu \tag{206}$$

($k = 1, \cdots, n$). Thus, combining the general result of Eq. (206) with the result of Eq. (203) we obtain that the conditional mean occupancy of site k, conditioned on the information that the site is not empty, is given by

$$\langle X_k | X_k > 0 \rangle \simeq \frac{\lambda}{\mu}\sqrt{\pi k} \ . \tag{207}$$

The power-law asymptotics of Eqs. (203) and (207) imply that the stationary occupation of 'downstream' sites (large k) exhibits large fluctuations. On the one hand, a downstream site is rarely occupied: $\Pr(X_k > 0) \simeq 1/\sqrt{\pi k}$. On the other hand, when a downstream site is occupied then its conditional mean is dramatically larger than its mean — the former being of order $O(\sqrt{k})$, while the latter being of order $O(1)$.

Fluctuations. A square root law of fluctuation, in which the variance in the occupancy of site k grows like $\sqrt{k}$, was numerically observed in Chapter 4. Here we prove that

$$\sigma^2(X_k) \simeq \frac{4\lambda^2}{\mu^2}\sqrt{\frac{k}{\pi}} \ . \tag{208}$$

Equation (208) is obtained by substituting Eq. (204) into the second moment $\langle X_k^2 \rangle = \sum_{l=1}^{\infty} l^2 \Pr(X_k = l)$, approximating the second moment by a corresponding integral, and noting that $\sigma^2(X_k) \simeq \langle X_k^2 \rangle$ (as the mean $\langle X_k \rangle$ is constant in k).

Inter-exit times. Consider the times at which particle clusters exit site k, and let T_k denote the time elapsing between two such consecutive exit events at steady state. Here we prove that the probability density of the scaled inter-exit time $T_k/\sqrt{\pi k}$ is asymptotically governed by the Rayleigh distribution

$$P_{T_k/\sqrt{\pi k}}(t) \simeq \frac{\pi t}{2} \exp\left(-\pi t^2/4\right) \tag{209}$$

($t > 0$), as previously anticipated by Monte-Carlo simulations.

Incremental load. The ASIP's overall load is the total number of particles present in the lattice at steady state. The steady state distribution of the overall load was comprehensively analyzed in Chapter 5. Generalizing the concept of the overall load we consider a 'lattice interval', contained within the ASIP lattice, which starts at site k and consists out of m consecutive sites: $\{k, k+1, \cdots, k+m-1\}$ ($k, m = 1, 2, 3, \cdots$). The ASIP's incremental load corresponding to this lattice interval at steady state is given by

$$L\,(k, m) = \sum_{i=k}^{k+m-1} X_i \ . \tag{210}$$

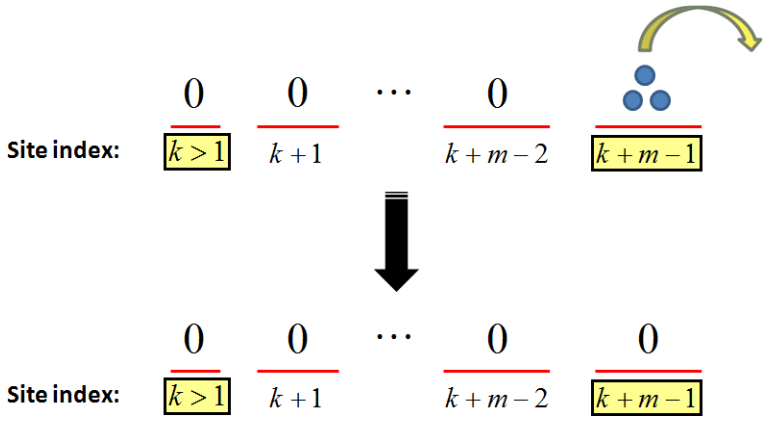

Figure 21: The non-empty interval $\{k, \ldots, k+m-1\}$ becomes empty if, and only if, all interval sites other than site $k+m-1$ are empty and the particles that occupy site $k+m-1$ hop to site $k+m$.

Clearly, the number of particles occupying site k, $L(k,1)$, and the overall load, $L(1,n)$, are both special cases of the incremental load $L(k,m)$. The main result of this chapter is an exact closed-form expression for the distribution of the incremental load

$$P_l(k,m) \equiv \Pr\big(L(k,m) = l\big)\,, \tag{211}$$

$(l = 0, 1, 2, \cdots)$. This expression, presented in Eq. (261), is given in terms of the entries of Catalan's trapezoids (see Chapter 7).

8.2 The ASIP as a coagulation model

As discussed in Chapter 3, coagulation models similar to the ASIP have been analyzed successfully using the empty-interval method and its generalization to non-empty intervals. In this method, one studies the steady state distribution of the incremental load defined in Eq. (210), and the time evolution of its associated time-dependent counterpart

$$L\,(t;k,m) = \sum_{i=k}^{k+m-1} X_i(t)\;, \tag{212}$$

where $X_i(t)$ denotes the number of particles present in site i at time t $(t \geq 0)$. In this section we review the method and show how it is applied to the analysis of the ASIP.

We begin with the probability that the lattice interval $\{k, k+1, \cdots, k+m-1\}$ is empty at time t:

$$P_0(t;k,m) \equiv \Pr\big(L(t;k,m) = 0\big). \tag{213}$$

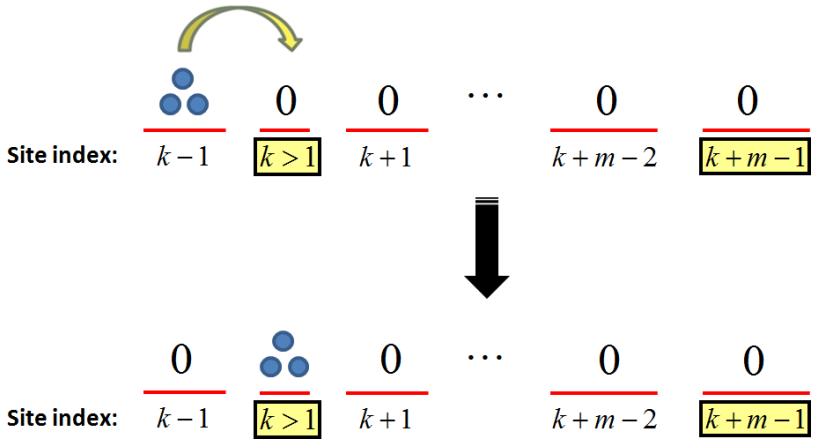

Figure 22: The empty interval $\{k, \ldots, k+m-1\}$ becomes non-empty if, and only if, site $k-1$ is occupied and the particles that occupy it hop to site k.

The empty-interval method is based on the fact that it is possible to write a closed-form evolution equation for the probabilities $P_0(t; k, m)$ as follows.

Consider a homogeneous ASIP. By rescaling time, the homogeneous gate opening rate and the particle arrival rate can be normalized to $\mu \to 1$ and $\lambda \to \lambda/\mu$ correspondingly. Accordingly, from this point onward we will assume, without loss of generality, that $\mu = 1$ and that λ is measured in units of the gate opening rate. For $k > 1$ and $m > 1$, the probability $P_0(t; k, m)$ evolves according to the equation

$$\begin{aligned} \tfrac{\partial}{\partial t} P_0(t; k, m) &= [P_0(t; k, m-1) - P_0(t; k, m)] \\ &- \big[P_0(t; k, m) - P_0(t; k-1, m+1)\big] \,. \end{aligned} \tag{214}$$

The term $P_0(t; k, m-1) - P_0(t; k, m)$ appearing on the right-hand side of Eq. (214) manifests the probability that sites $\{k, k+1, \cdots, k+m-2\}$ are empty and site $k+m-1$ is occupied, in which case the particle cluster at site $k+m-1$ might hop (with rate 1) to site $k+m$ and thus leave the interval $\{k, \cdots, k+m-1\}$ empty, as illustrated in Figure 21. Similarly, the term $P_0(t; k, m) - P_0(t; k-1, m+1)$ appearing on the right-hand side of Eq. (214) manifests the probability that sites $\{k, k+1, \cdots, k+m-1\}$ are empty and site $k-1$ is occupied, in which case the particle cluster at site $k-1$ might hop to site k (with rate 1), thus rendering the interval $\{k, k+1, \cdots, k+m-1\}$ non-empty as illustrated in Figure 22.

Eq. (214) remains valid for $m = 1$ and $k > 1$ provided that we impose the boundary condition

$$P_0(t; k, 0) \equiv 1 \,, \tag{215}$$

i.e., degenerate intervals (which contain no sites) are by convention always

empty. For $k = 1$ and $m \geq 1$ the evolution is given by

$$\frac{\partial}{\partial t} P_0(t; 1, m) = [P_0(t; 1, m-1) - P_0(t; 1, m)] - \lambda P_0(t; 1, m)\,. \tag{216}$$

The term $P_0(t; 1, m-1) - P_0(t; 1, m)$ appearing on the right-hand side of Eq. (216) manifests the probability that sites $\{1, 2, \cdots, m-1\}$ are empty and site m is occupied, in which case the particle cluster at site m might hop (with rate 1) to site $m+1$ and thus leave the interval $\{1, \cdots, m\}$ empty. Also, $P_0(t; 1, m)$ is the probability that the interval $\{1, \cdots, m\}$ is empty, in which case a particle might arrive to site 1 (with rate λ), thus rendering the interval $\{1, \cdots, m\}$ non-empty.

The empty-interval method can be generalized to capture the evolution of the probability $P_l(t; k, m)$ that there are exactly l particles at sites $\{k, k+1, \cdots, k+m-1\}$ at time t [55, 54]:

$$P_l(t; k, m) \equiv \Pr\big(L(t; k, m) = l\big). \tag{217}$$

The empty-interval probabilities $P_0(t; k, m)$ are hence a special case of $P_l(t; k, m)$ with $l = 0$. The counterparts of Eqs. (214)–(216) are as follows (see Appendices 8.8.1 and 8.8.2 for the derivations). For $k > 1$ and $m > 1$ the evolution is given by

$$\begin{aligned} &\tfrac{\partial}{\partial t} P_l(t; k, m) = \\ &+\Big[P_l(t; k, m-1) - 2P_l(t; k, m) + P_l(t; k, m+1)\Big] \\ &-\Big[P_l(t; k, m+1) - P_l(t; k-1, m+1)\Big]\,. \end{aligned} \tag{218}$$

Equation (218) remains valid for $m = 1$ and $k > 1$ provided that we impose the boundary condition

$$P_l(t; k, 0) = \delta_{l,0}\,. \tag{219}$$

where $\delta_{l,0}$ is the Kronecker delta symbol. Note that, remarkably, Eqs. (218) for $P_l(t; k, m)$ do not couple different values of l. A coupling enters only through the boundary condition $P_l(t; 1, m)$ $(m \geq 1)$ whose time evolution is given by

$$\begin{aligned} &\tfrac{\partial}{\partial t} P_l(t; 1, m) = \\ &+\Big[P_l(t; 1, m-1) - P_l(t; 1, m)\Big] \\ &-\lambda\Big[P_l(t; 1, m) - P_{l-1}(t; 1, m)\Big]\,. \end{aligned} \tag{220}$$

Note that setting $l = 0$ in Eqs. (218)–(220), while taking into account that the probability to observe a negative number of particles is zero by definition, indeed yields Eqs. (214)–(216).

8.3 Continuum limits of the steady-state equations

The main result of this chapter is an exact expression for the steady-state solution of Eqs. (214)–(220). Before presenting and deriving this exact solution (see Sections 8.5 and 8.6) we provide in the current section a derivation of the asymptotic scaling forms that this solution attains for large values of k, i.e., for lattice intervals located far away downstream. Some of the asymptotic results presented in this section have been obtained before in [56] using Laplace transform methods. Here we present an alternative "real-space" derivation, which yields new physical insight into the solutions and highlights their universal nature.

The asymptotic analysis of Eqs. (214)–(220) is based on the following continuum-limit assumption: if the steady state probability $P_l(k,m)$ changes slowly as a function of the variables k and m, then this discrete function may be approximated by one which is continuous both in k and m. Thus, one can expand to leading order all terms in the equation around $P_l(k,m)$. In this continuum limit, the discrete Laplacian in the first square brackets of Eq. (218) approximately equals a continuous Laplacian, and similarly the second square brackets is approximately $\frac{\partial}{\partial k}P_l(k,m)$. Therefore, in the steady-state, where the left-hand side of Eq. (218) vanishes, one finds that $P_l(k,m)$ satisfies a diffusion equation where the site number k plays the role of time:

$$\frac{\partial}{\partial k}P_l(k,m) = \frac{\partial^2}{\partial m^2}P_l(k,m)\,. \tag{221}$$

This continuum approximation will be shown *a-posteriori* to be valid when $k \gg m$.

Equation (221) should be solved with the appropriate boundary conditions in "space" (i.e., in m) and "time" (i.e., in k). The spatial ($m = 0$) boundary condition of Eq. (221) is given in Eq. (219), $P_l(k,0) = \delta_{l,0}$. The temporal ($k = 1$) initial condition is the steady state solution of Eq. (220), which was found to be (see Section 5.6):

$$P_l(1,m) = \binom{l+m-1}{l}\left(\frac{1}{1+\lambda}\right)^m\left(\frac{\lambda}{1+\lambda}\right)^l. \tag{222}$$

Before proceeding with the study of Eq. (221), let us discuss its relation with the behavior of an ASIP on a ring. Unlike the open boundary ASIP on which we focus, on a ring the steady-state behavior of the model is trivial: a single occupied site circulates throughout the system uni-directionally. The relaxation to this steady-state, however, has an interesting scaling form which has been studied extensively in the context of coagulation models (see Chapter 3). In particular, it is known that in a spatially homogeneous ring, the probability to find l particles in an interval of m sites evolves (in a continuum limit) according to the diffusion equation (221) with k replaced by time. In other words, as one progresses from left to right along a stationary open-boundary ASIP, the probability to see empty or occupied intervals changes (in space) just like the temporal evolution of the corresponding probability on a ring. This mapping

between the two problems provides an interesting physical picture: it suggests that the open-boundary ASIP can be thought of as a sort of a "conveyor belt", along which the coagulation reaction proceeds. A single steady-state snapshot of the open-boundary ASIP is, in this sense, similar to the entire temporal evolution of the coagulation model on a ring.

It is well known that the diffusion equation on an infinite line has, at times which are large compared with (the square of) the spatial extent of the initial condition, a scaling form of a spreading Gaussian. Having arrived at the diffusion equation (221), it is not too surprising that a similar scaling solution is found for it at large k. This solution, however, is not Gaussian, due to the boundary condition (219) which is either a source at the origin when $l = 0$ or a sink when $l \geq 1$. In Subsections 8.3.1 and 8.3.2 below, we separately describe and derive the scaling solutions for these two cases. A third, somewhat more subtle, scaling solution is found when considering the joint limit of $l \sim \sqrt{k} \gg 1$. In this case, k is not large enough in comparison with the spatial extent of the "initial condition" (222) in order for the usual scaling of the diffusion equation to apply. Nonetheless, $P_l(k,m)$ is found to have a universal scaling form in the variable $l/\sqrt{k}$. This scaling form is discussed in Subsection 8.3.3. The universality of the obtained scaling forms and the conditions under which the continuum approximation is valid are discussed in Subsection 8.3.4.

8.3.1 The case of $l = 0$

As with the usual (probability conserving) diffusion in its late stages, the large k solution of Eq. (221) is given by a scaling form. This form can be found by substituting the ansatz

$$P_l(k,m) = k^{-\beta} f\left(\frac{m}{\sqrt{k}}\right) \tag{223}$$

in Eq. (221), yielding the ordinary differential equation

$$f''(u) + \frac{u}{2} f'(u) + \beta f(u) = 0 \tag{224}$$

for the scaling function $f(u)$, where $u = m/\sqrt{k}$ is the corresponding scaling variable.

In the case of $l = 0$ (i.e., the probability to see empty intervals), the boundary condition $P_0(k,0) = 1$ implies that $\beta = 0$ and $f(0) = 1$. The solution of Eq. (224) with this boundary condition is given by $f(u) = 1 + C\,\mathrm{erf}(u/2)$ where C is an integration constant, and $\mathrm{erf}(x) \equiv 2/\sqrt{\pi} \int_0^x \exp(-y^2) dy$ is the error function. For large u this solution approaches $1 + C$. Since $\lim_{m\to\infty} P_0(k,m) \to 0$ (i.e., there is a vanishing probability that all sites from k onwards are empty), the constant C must equal -1, yielding the scaling solution $f(u) = \mathrm{erfc}(u/2)$, i.e.,

$$P_0(k,m) \simeq \mathrm{erfc}\Big(\frac{m}{2\sqrt{k}}\Big), \qquad (m \ll k) \tag{225}$$

where erfc is the complementary error function defined as $\mathrm{erfc}(x) \equiv 1 - \mathrm{erf}(x)$. Here and in the next two subsections we indicate in brackets the limiting regime in which the obtained scaling solutions are valid. These are explained below in Subsection 8.3.4.

8.3.2 The case of $1 \leq l \ll \sqrt{k}$

When $l \geq 1$, Eq. (221) should be solved under the absorbing boundary condition $P_l(k,0) = 0$, which by use of Eq. (223) implies that $f(0) = 0$. The corresponding solution of Eq. (224) is

$$f(u) = C\, u\, {}_1F_1(\beta + 1/2; 3/2; -u^2/4), \tag{226}$$

where C is once again an integration constant and ${}_1F_1(a;b;z)$ is the Kummer hypergeometric function. The values of β and C can be determined by using the fact that the quantity $\Lambda = \int_0^\infty m P_l(k,m) dm$ is conserved by the diffusion equation (221) with an absorbing boundary condition, i.e., it can be shown that $d\Lambda/dk = 0$ [81]. The discrete counterpart of this conservation law, which results from Eq. (218), states that

$$\Lambda_l \equiv \sum_{m=1}^{\infty} (m-1) P_l(k,m) \tag{227}$$

is independent of k in the steady state. For the scaling solution given by the combination of Eqs. (223) and (226), $\Lambda_l \simeq k^{1-\beta} \int u f(u) du = k^{1-\beta}\sqrt{4\pi} C$, and we therefore find that $\beta = 1$, for which $f(u) = Cu\exp(-u^2/4)$ [82], and $C = \Lambda_l/\sqrt{4\pi}$, i.e.,

$$P_l(k,m) \simeq \frac{\Lambda_l m}{\sqrt{4\pi} k^{3/2}} e^{-\frac{m^2}{4k}} \qquad (1 \leq l \ll \sqrt{k};\ m \ll k). \tag{228}$$

The value of Λ_l is found from the initial condition (222) to be

$$\Lambda_l = \sum_{m=1}^{\infty} (m-1) P_l(1,m) = (l+1)/\lambda^2. \tag{229}$$

To see this, note that up to a multiplication by λ^{-1}, Eq. (222) is the probability mass function of a sum of $l+1$ independent geometric random variables with mean λ^{-1}.

Note that the scaling form (228) is valid only in the asymptotic regime when the diffusive length $\sqrt{k}$ is much larger than the spatial spread of the initial condition, which in our case is of the same order of Λ_l. In other words, for any fixed $l \geq 1$, Eq. (228) is a good approximation at "times" where $\sqrt{k} \gg l$. In the next subsection we examine what happens at "times" $\sqrt{k} \sim l$ which are not large enough for the initial condition to be washed out by the diffusion.

8.3.3 The case of $l \sim \sqrt{k}$

When $l \sim \sqrt{k}$ and k is not large enough for the diffusion to reach its asymptotic scaling regime, there seems to be no a-priori reason to expect a scaling solution to Eq. (221). However, a closer inspection of the initial condition (222) reveals that such a scaling solution does exist and, surprisingly, is also universal. We now derive this scaling solution; its universality is discussed in the next subsection.

The key observation now is that the dependence on the number of particles l enters only through the initial condition of Eq. (222), which in the limit we study, and as a function of m, is narrowly centered around $m \simeq l/\lambda$. This once again follows from the fact that the initial condition of Eq. (222) is proportional to the probability mass function of a sum of $l+1$ independent geometric random variables with mean λ^{-1}. Therefore, according to the central limit theorem, the distribution of this sum can be approximated, when $l \to \infty$, by a Gaussian distribution whose mean is given by $\langle m \rangle = (l+1)/\lambda \simeq l/\lambda$. Recalling that the standard deviation scales as $\sqrt{l}$, and is therefore negligible with respect to the mean, we can further approximate the Gaussian probability density function by a Dirac δ function, i.e.,

$$P_l(1,m) \simeq \lambda^{-1}\,\delta(m - l/\lambda). \tag{230}$$

The solution of the diffusion equation (221) with an absorbing boundary at the origin and the initial condition (230) is found (e.g., by the method of images [83]) to be

$$P_l(k,m) \simeq \frac{1}{\sqrt{4\pi\lambda^2 k}}\left[e^{-\frac{(m-l/\lambda)^2}{4k}} - e^{-\frac{(m+l/\lambda)^2}{4k}}\right]$$

$$(1 \ll l \ll k;\; m \ll k). \tag{231}$$

Equation (231) is a joint scaling solution in the scaling variables $m/\sqrt{k}$ and $l/\sqrt{k}$. If one is further interested in the limit of $m \ll l$, one may expand and obtain to leading order a "thermal dipole"

$$P_l(k,m) \simeq \frac{ml}{\sqrt{4\pi}\lambda^2 k^{3/2}} e^{-\frac{l^2}{4\lambda^2 k}} \qquad (m \ll l \ll k). \tag{232}$$

Note that, as explained below, Eqs. (231) and (232) are valid not only at the scale of $l \sim \sqrt{k}$, but in fact for all $1 \ll l \ll k$.

8.3.4 Remarks on the scaling solutions

In this subsection we remark on the limits of validity of the scaling solutions obtained above, and discuss their universality.

The validity of the scaling solutions obtained in the previous sections relies on the continuum approximation of the exact (discrete) Eq. (218) by the continuous

Eq. (221). A straightforward calculation shows that the solutions (225), (228), (231), and (232) satisfy

$$P_l(k,m+1) - P_l(k-1,m+1) = \frac{\partial}{\partial k} P_l(k,m) \left[1 + O\Big(\frac{m}{k}, \frac{l}{k}\Big) \right]. \qquad (233)$$

and similarly for the discrete m-Laplacian. Therefore, the continuum approximation is valid as long as $m, l \ll k$. Note in particular that the continuum limit does not require m to be large, and thus the results are valid even for $m = 1$.

An important feature of the scaling solutions (225), (228), (231), and (232) is their *universality* with respect to the details of the how particles arrive at the first site: while the arrival process dictates the distribution of $L(1,m)$, i.e., the initial condition $P_l(1,m)$, the scaling solutions are rather insensitive to it. In other words, one may say that the arrival process which feeds particles into the ASIP "conveyor belt" does not affect the load statistics far away downstream. As discussed shortly, universality breaks down for some exotic initial conditions with fat tails, but is otherwise expected to hold for a rather large class of arrival processes.

For the scaling solutions (225) and (228), the origin of universality is easily understood from the diffusion picture of Eq. (221): it is well known that solutions of the diffusion equation converge at late times to scaling functions that are independent of the initial condition (as long as the tail of the initial condition decays rapidly enough) [81]. We now note that $P_0(1,m) \leq P_0(2,m-1) \leq 1/2^{m-1}$ for any arrival process, as can be clearly seen by considering a limiting scenario in which the arrival process is such that the first site is always occupied. Hence, the initial condition $P_0(1,m)$ decays (at least) exponentially fast in m and the pathological case of heavy tails is excluded. As a result, Eq. (225) is not only independent of λ in the case of Poissonian arrivals but also completely insensitive to nature of the arrival process altogether.

Equation (228) is also universal, except for the prefactor Λ_l given by Eq. (229). This prefactor (and only it) depends on the details of the arrival process, and is thus non universal. However, when $l \gg 1$ and for initial conditions which can be approximated by Eq. (230) (see discussion shortly), the prefactor attains the universal form $\Lambda_l \simeq l/\lambda^2$. This form still "remembers" the mean arrival rate λ, but is otherwise independent of the arrival process. Its dependence on λ is both mathematically unavoidable, due to the conservation of Λ_l, and physically reasonable, as the mean number of particles per site in Eq. (206) depends on λ. The universality of Eq. (228) breaks down for fat-tailed $P_l(1,m)$ for which Λ_l diverges.

The universality of Eqs. (231) and (232) has a somewhat more subtle origin. As explained above, these scaling forms are valid even though k is *not* large enough to "wash out" the initial condition $P_l(1,m)$. Rather, they emerge exactly when the diffusive length $\sqrt{k}$ is of the order of the initial length scale $\sum_m m P_l(1,m) \sim l$. The validity of these scaling functions rests on the approximation in Eq. (230), which itself is a result of the central limit theorem. Therefore, the scaling forms (231) and (232) hold whenever the arrival process is

such that $L(1,m)$ lends itself to one of the many extensions and generalizations of the central limit theorem. This universality is demonstrated by a specific exactly-solvable example in Appendix 8.8.3.

The scaling forms (231) and (232) will hold even when the central limit theorem breaks down, and as long as the standard deviation in $L(1,m)$ is negligible with respect to its mean in the limit of $m \to \infty$. When this is the case, the distribution of $L(1,m)$ is sharply peaked around its mean thus asserting the existence of an approximation of the type appearing in Eq. (230). The basin of attraction for this type of behavior is very large. Indeed, for a general arrival process, Little's law [9] asserts that $\langle L(1,m) \rangle = \bar{\lambda} m$, where $\bar{\lambda}$ is the effective arrival rate (long term average of the number of particles arriving per unit time) and m is the average time a particle spends in the system. On the other hand, fluctuations in $L(1,m)$ are only caused by arrivals to the first site and departures from the last site. And so, given the universality of Eq. (225), if the typical fluctuation due to an arrival event is finite and when m is large, fluctuations in $L(1,m)$ will be dominated by departure events. Hence, the standard deviation in $L(1,m)$ will be of order $\sqrt{m}$ and, most importantly, negligible with respect to the mean.

8.4 Implications of the inter-particle distribution function

In this section we use the results of Section 8.3 to derive the scaling properties of the ASIP which were presented in Section 8.1.

Occupation probabilities. We begin by examining the probability that a site is occupied. Substituting $m = 1$ in Eq. (225) and expanding to first order in k, we recover Eq. (203). The occupation-number distribution of a single site, $P_l(k,1)$, is found by substituting $m = 1$ in Eq. (232). Recalling that we have rescaled time such the $\mu = 1$, we recover the scaling form reported in Eqs. (204) and (205). In fact, combining (228) with (232) we may write a uniform approximation which is asymptotically exact for all $l \geq 1$ in the limit of $k \gg 1$:

$$P_l(k,1) \simeq \frac{\Lambda_l}{\sqrt{4\pi} k^{3/2}} e^{-\frac{l^2}{4\lambda^2 k}}, \tag{234}$$

where Λ_l is given in (229). An interesting picture emerges from the above-mentioned results. Downstream sites with $k \gg 1$ are mostly empty. However, conditioned on being occupied, their occupation is typically of the order of $\sqrt{k}$ [see Eq. (207)], and in fact its distribution has the scaling form of Eq. (234). Below, in Section 8.6, we derive an exact expression for this occupation probability which is correct even for small k.

Inter-particle distance probability. Another quantity of interest is the inter-particle distance probability $Q(k,m)$, which is defined as the conditional probability that the next occupied site after site k is site $k+m$ given that site k itself is occupied. The scaling solutions found in Section 8.3 allow us to calculate $Q(k,m)$. To do so, we first examine the unconditional probability $(1 - P_0(k,1))\, Q(k,m)$ that sites k and $k+m$ are both occupied and the $m-1$

sites in between the two are empty. This probability is given by

$$\begin{aligned}(1-P_0(k,1))\,Q(k,m) &= P_0(k+1,m-1)\\ &-[P_0(k+1,m)-P_0(k,m+1)]\\ &-[P_0(k,m)-P_0(k,m+1)]-P_0(k,m+1).\end{aligned} \tag{235}$$

The first term in Eq. (235) is the probability that sites $\{k+1,\cdots,k+m-1\}$ are empty. From this probability one must subtract: (i) the probability that these sites are empty, site k is occupied and site $k+m$ is empty (the second term, in square brackets); (ii) the probability that these sites are empty, site k is empty and site $k+m$ is occupied (the third term, in square brackets); (iii) the probability that all $m+1$ sites from k to $k+m$ are empty (the last term). Rearranging and passing, as before, to a continuum limit yields

$$\begin{aligned}&(1-P_0(k,1))\,Q(k,m) =\\ &+[P_0(k,m+1)-2P_0(k,m)+P_0(k,m-1)]\\ &-[P_0(k+1,m)-P_0(k,m)-P_0(k+1,m-1)+P_0(k,m-1)]\\ &\simeq\left(\tfrac{\partial^2}{\partial m^2}-\tfrac{\partial}{\partial m\partial k}\right)P_0(k,m)|_{k,m}\,.\end{aligned} \tag{236}$$

Substituting Eq. (225) we see that the $\partial^2/\partial m^2$ term dominates in the large k limit, and we obtain

$$Q(k,m)\simeq\frac{f''(u)}{k\,(1-P_0(k,1))}=\frac{ue^{-u^2/4}}{2\sqrt{k}}, \tag{237}$$

where once again $f(u)=\mathrm{erfc}(u/2)$ and $u=m/\sqrt{k}$.

Inter-exit times. For an ASIP in steady state let the random variable T_k denote the time elapsing between two consecutive time epochs at which particles exit site k. Equation (237) allows us to evaluate the typical order of magnitude of T_k in the limit of $k\gg 1$. Indeed, given that site k is occupied, it will take (on average) a single time unit for particles to hop out of it — resulting in a first exit event. On the other hand, we know that $Q(k-m,m)$ is the probability that $k-m$ is the nearest occupied site in the upstream direction. The average distance to the nearest occupied site is hence

$$\begin{aligned}&\textstyle\sum_{m=1}^{k-1} mQ(k-m,m)\simeq\\ &\sqrt{k}\int_0^{\sqrt{k}}\frac{u^2e^{-u^2/4(1-u/\sqrt{k})}}{2\sqrt{1-u/\sqrt{k}}}du\simeq\sqrt{\pi k}\,.\end{aligned} \tag{238}$$

Thus, $1+\sqrt{\pi k}$ sites on average are to be traversed at an average 'speed' of one site per unit time for the second exit event to occur. When k is large, T_k is clearly dominated by this traversal time. The error incurred by neglecting the time awaited till the occurrence of the first exit event is negligible and we may safely conclude that $\langle T_k \rangle / \sqrt{\pi k} \simeq 1$.

We can further go on and compute the asymptotic distribution of the inter-exit time. To see how, note that in the limit of $k \gg 1$, the reasoning given above asserts that the probability density of the random variable T_k may be approximated by

$$P_{T_k}(t) \simeq \sum_{m=1}^{k-1} \frac{t^m e^{-t}}{m!} Q(k-m,m) \tag{239}$$

where $t^m e^{-t}/m!$ is the probability density for the traversal time of $m+1$ sites. In Appendix 8.8.4 we show that the sum in (239) can be evaluated using a saddle point approximation to yield Eq. (209).

8.5 Incremental Load Analysis

The analysis conducted so far was based on a continuum limit approximation of Eq. (218) at steady state. Using this approach we were able to analyze homogenous ASIPs and obtain an asymptotic solution for the probabilities $P_l(k,m)$ in the limit $k \gg m,l$. We now set forth to obtain an *exact* solution for this problem. In order to demonstrate the general applicability of the approach described hereinafter we develop it in the context of general ASIPs (not necessarily homogeneous). Setting off from the stochastic law of motion of the incremental load, we go on to derive the boundary value problem which governs its steady state distribution. An algorithm for the solution of this problem is presented along with iterative schemes for the computation of occupation probabilities and factorial moments. In the next section we return to the case of homogeneous ASIPs.

8.5.1 The Incremental Load

In this section we revisit the notion of incremental load, which generalizes the notion of overall load. In what follows we consider an infinite lattice with countably many sites, and analyze the ASIP's incremental load in detail. We consider the lattice interval starting at site k and consisting of m sites — $\{k, k+1, \cdots, k+m-1\}$ $(k,m = 1,2,3,\cdots)$ — and remind the reader that the ASIP's incremental loads $L(k,m)$ and $L(t;k,m)$ are given by Eqs. (210) and (212), respectively.

Throughout this section we shall employ the natural boundary conditions $L(t;k,0) = 0$ and $L(k,0) = 0$. The Probability Generating Functions (PGFs) of the incremental loads $L(k,m)$ and $L(t;k,m)$ are given, respectively, by

$$G(z;k,m) = \left\langle z^{L(k,m)} \right\rangle , \tag{240}$$

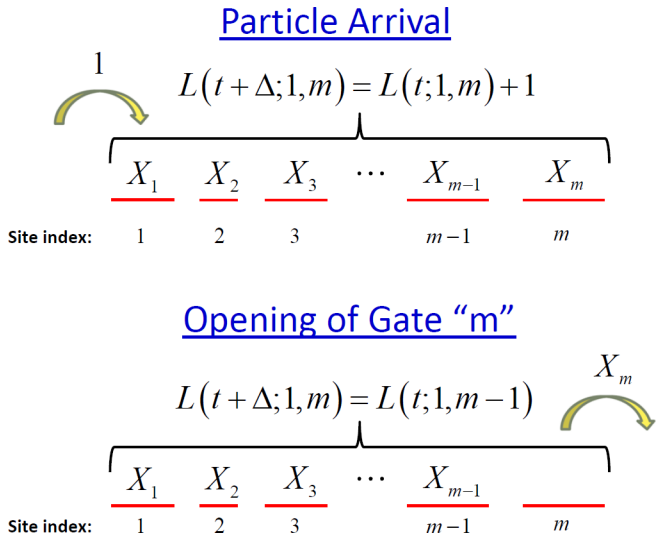

Figure 23: During the time interval $(t, t+\triangle)$, the incremental load $L(t;1,m)$ can change either due to the arrival of a particle to the first site, or due to the opening of gate m.

and

$$G(t,z;k,m) = \left\langle z^{L(t;k,m)} \right\rangle , \tag{241}$$

($|z| \leq 1$). Note that the boundary conditions $L(t;k,0) = 0$ and $L(k,0) = 0$ imply, respectively, the following PGF boundary conditions:

$$G(z;k,0) = 1 \quad \text{and} \quad G(t,z;k,0) = 1. \tag{242}$$

8.5.2 The Case of $k = 1$

In this subsection we analyze the special case of lattice intervals initiating at the first lattice site $k = 1$. This special case yields the overall load which was analyzed in Chapter 5 via the ASIP's multidimensional PGF. Here we analyze this special case via the method of incremental loads. This serves to illustrate the method which will later be used to derive new results for $k > 1$.

Consider the lattice interval starting at site 1 and consisting of m sites, and observe its incremental load at times t and $t' = t + \Delta$ (where $\Delta \to 0$). During the time interval (t, t') exactly two events, illustrated in Figure 23, can change the incremental load. One event is the arrival of a particle to the lattice — in which case the arriving particle enters the first site and hence $L(t';1,m) = L(t;1,m) + 1$; this event occurs with probability $\lambda\Delta + o(\Delta)$. The other event is the opening of gate m — in which case all particles present in site m transit to site $m+1$ and hence $L(t';1,m) = L(t;1,m-1)$; this event occurs with probability $\mu_m\Delta + o(\Delta)$. Note that the boundary condition $L(t;1,0) = 0$ indeed fits in naturally. If neither of these two events take place — a scenario

occurring with probability $1-(\lambda+\mu_m)\,\Delta+o(\Delta)$ — then the incremental load is left unchanged: $L\,(t';1,m)\;=\;L\,(t;1,m)$. Thus, the stochastic connection between the incremental loads $L\,(t;1,m)$ and $L\,(t';1,m)$ is given by

$$L\,(t';1,m)= \begin{cases} L\,(t;1,m)+1 & \text{w.p.}\quad \lambda\Delta+o(\Delta)\ , \\ L\,(t;1,m-1) & \text{w.p.}\quad \mu_m\Delta+o(\Delta)\ , \\ L\,(t;1,m) & \text{w.p.}\quad 1-(\lambda+\mu_m)\,\Delta+o(\Delta)\ , \end{cases} \tag{243}$$

and we note that "w.p." is used here as a short hand for the term "with probability".

Shifting from the incremental loads $L\,(t;1,m)$ and $L\,(t';1,m)$ to their respective PGFs, Eq. (243) yields the following PGF dynamics

$$\begin{aligned} &\tfrac{\partial}{\partial t}G\,(t,z;1,m)= \\ &\left[\lambda\,(z-1)-\mu_m\right]G\,(t,z;1,m)+\mu_m G\,(t,z;1,m-1)\ . \end{aligned} \tag{244}$$

The derivation of Eq. (244) is given in Appendix 8.8.5. At steady state the time-dependence vanishes, and the differential equation (244) reduces to the steady-state equation

$$G\,(z;1,m)=\frac{\mu_m}{\mu_m+\lambda\,(1-z)}G\,(z;1,m-1)\ . \tag{245}$$

A straightforward iterative solution of Eq. (245), using the PGF boundary condition $G\,(z;1,0)=1$, yields the following explicit form for the PGF of the incremental load at steady state

$$G\,(z;1,m)=\prod_{i=1}^{m}\frac{1}{1+\frac{\lambda}{\mu_i}\,(1-z)}\,. \tag{246}$$

Note that the terms λ/μ_i appearing in Eq. (246) are the ratios of the particles' inflow rate to the gates' opening rates, as well as the mean occupancies at steady state ($\lambda/\mu_i=\langle X_i\rangle$) (see Chapter 5).

Eq. (246) has several important implications. Firstly, Eq. (246) implies that at steady state the overall load $L\,(1,1)$ of a *single-site* ASIP ($n=1$) follows a *geometric distribution*. Indeed, setting $m=1$ in Eq. (246) yields the PGF of the following geometric probability distribution: $\Pr\left(L\,(1,1)=l\right)=(1-p_1)^l p_1$ ($l=0,1,2\cdots$), where $p_1=\mu_1/\left(\mu_1+\lambda\right)$. Secondly, the *product-form* structure of Eq. (246) implies that at steady state the overall load $L\,(1,m)$ admits the stochastic representation

$$L(1,m)=\sum_{i=1}^{m}G_i\ , \tag{247}$$

where $\{G_1, \cdots, G_m\}$ is a sequence of independent geometrically-distributed random variables: $\Pr(G_i = l) = (1-p_i)^l p_i$ $(l = 0, 1, 2, \cdots)$, with $p_i = \mu_i/(\mu_i + \lambda)$ $(i = 1, \cdots, m)$. The overall load $L(1, m)$ is hence equal, in law, to the sum of the overall loads of m *independent single-site* ASIPs with respective parameters $(\lambda, \mu_1), \cdots, (\lambda, \mu_m)$. Thus, the distribution of the overall load $L(1, m)$ is given by

$$P_l(1,m) = \Pr(L(1,m) = l) = \sum_{l_1, \cdots, l_m} \left(\prod_{i=1}^{m} p_i (1-p_i)^{l_i} \right) \delta\left(l - \textstyle\sum_i l_i\right) , \tag{248}$$

where the Dirac δ function guarantees that $\sum_i l_i = l$. Thirdly, setting $z = 0$ in Eq. (246) (or $l = 0$ in Eq. (248)) yields the probability that the lattice interval $\{1, \cdots, m\}$ is empty

$$P_0(1,m) = \Pr(L(1,m) = 0) = \prod_{i=1}^{m} \frac{\mu_i}{\mu_i + \lambda} . \tag{249}$$

8.5.3 The Case $k > 1$

In this subsection we analyze the general case of lattice intervals initiating at an arbitrary lattice site $k > 1$. While the special case $k = 1$ could be analyzed via the ASIP's joint PGF, an analogous analysis of the general case $k > 1$ via this method is prohibitively hard. However, as we shall now demonstrate, the analysis of the general case $k > 1$ is well attainable following an approach parallel to the one applied in the previous subsection.

Consider the lattice interval starting at site k $(k > 1)$ and consisting of m sites, and observe its incremental load at times t and $t' = t + \Delta$ (where $\Delta \to 0$). During the time interval (t, t') exactly two events, illustrated in Figure 24, can change the incremental load. One event is the opening of gate $k-1$ — in which case all particles present in site $k-1$ transit to site k and hence $L(t'; k, m) = L(t; k-1, m+1)$; this event occurs with probability $\mu_{k-1}\Delta + o(\Delta)$. The other event is the opening of gate $k+m-1$ — in which case all particles present in site $k+m-1$ transit to site $k+m$ and hence $L(t'; k, m) = L(t; k, m-1)$; this event occurs with probability $\mu_{k+m-1}\Delta + o(\Delta)$. As noted in Subsection 8.5.2, the boundary condition $L(t; k, 0) = 0$ fits in naturally. If neither of these two events take place — a scenario occurring with probability $1 - (\mu_{k-1} + \mu_{k+m-1})\Delta + o(\Delta)$ — then the incremental load is left unchanged: $L(t'; k, m) = L(t; k, m)$. Thus, the stochastic connection between the incremental loads $L(t; k, m)$ and $L(t'; k, m)$ is given by

$$L(t'; k, m) = \begin{cases} L(t; k-1, m+1) & \text{w.p.} \quad \mu_{k-1}\Delta , \\ L(t; k, m-1) & \text{w.p.} \quad \mu_{k+m-1}\Delta , \\ L(t; k, m) & \text{w.p.} \quad 1 - (\mu_{k-1} + \mu_{k+m-1})\Delta . \end{cases} \tag{250}$$

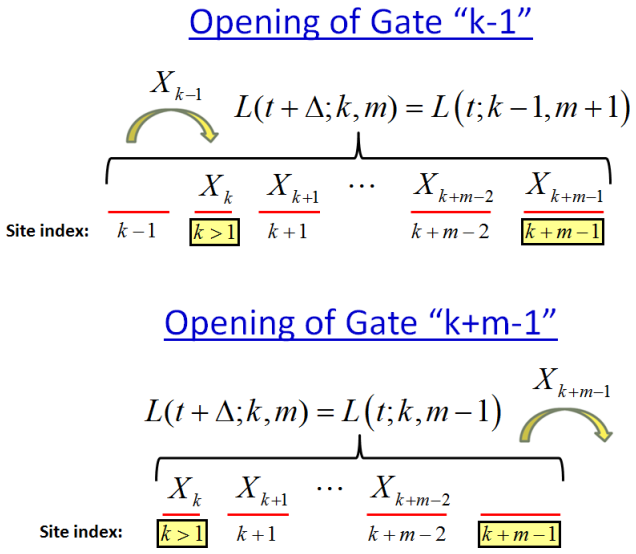

Figure 24: During the time interval $(t, t+\triangle)$, the incremental load $L(t;k,m)$ can change either due to the opening of gate $k-1$ or due to the opening of gate $k+m-1$.

Shifting from the incremental loads $L(t;k,m)$ and $L(t';k,m)$ to their respective PGFs, Eq. (250) yields the following PGF dynamics

$$\begin{aligned}\tfrac{\partial}{\partial t}G(t,z;k,m) &= -\left(\mu_{k-1}+\mu_{k+m-1}\right)G(t,z;k,m)\\ &+\mu_{k-1}G(t,z;k-1,m+1)+\mu_{k+m-1}G(t,z;k,m-1)\ .\end{aligned} \tag{251}$$

The derivation of Eq. (251) is given in Appendix 8.8.6. At steady state the time-dependence vanishes, and the differential equation (251) reduces to the steady-state equation

$$\begin{aligned}G(z;k,m) &= \tfrac{\mu_{k+m-1}}{\mu_{k-1}+\mu_{k+m-1}}G(z;k,m-1)\\ &+\tfrac{\mu_{k-1}}{\mu_{k-1}+\mu_{k+m-1}}G(z;k-1,m+1)\ .\end{aligned} \tag{252}$$

For any fixed z, Eq. (252) defines a two-dimensional boundary value problem for $G(z;k,m)$. The problem and an algorithm for its solution are illustrated in Figure 25.

Equation (252) can also be used to establish an explicit iterative scheme for the computation of the PGF $G(z;k,m)$ in terms of the PGFs $\{G(z,k-1,i)\}_{i=2,\ldots,m+1}$.

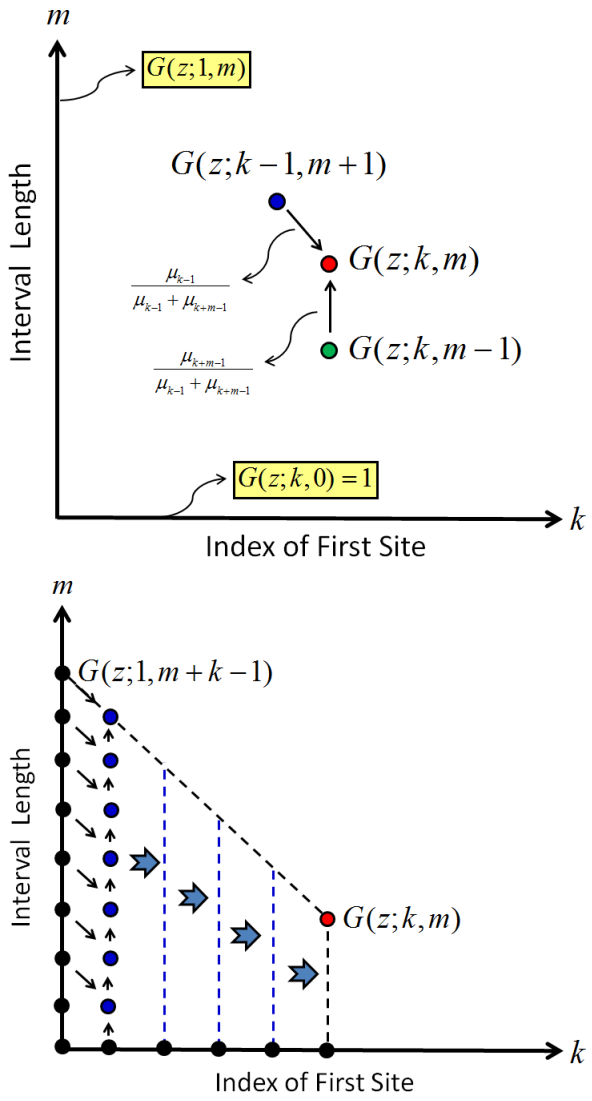

Figure 25: Top Panel: Equation (252) defines a boundary value problem for $G(z;k,m)$. The PGF $G(z;k,m)$ is determined by a weighted average of its southern and northwestern neighbors in the positive quadrant of the (k,m) plane. The boundary PGFs $G(z;k,0)$ and $G(z;1,m)$ are given by Eqs. (242) and (246) respectively. Bottom Panel: A three step algorithm can be used in order to solve the boundary value problem for the PGF $G(z;k,m)$: (i) start at the left boundary and solve for the column that stands to its right; (ii) treat the newly solved column as the new left boundary and iterate; (iii) stop at the k^{th} column and obtain the desired solution.

Specifically:

$$\begin{aligned} G\left(z;k,m\right) &= \Pi\left(k,m\right) \\ &+\Pi\left(k,m\right)\sum_{i=1}^{m}\frac{\mu_{k-1}}{\mu_{k-1}+\mu_{k+i-1}}\frac{G(z;k-1,i+1)}{\Pi(k,i)}\ , \end{aligned} \tag{253}$$

where

$$\Pi\left(k,m\right) = \prod_{j=1}^{m}\frac{\mu_{k+j-1}}{\mu_{k-1}+\mu_{k+j-1}}\ , \tag{254}$$

and where the boundary condition $G\left(z;1,m\right)$ is given by Eq. (246). The derivation of Eq. (253) is given in Appendix 8.8.7.

8.5.4 Occupation Probabilities and Factorial Moments

Based on the incremental-load results established hitherto, in this subsection we derive recursive equations for the occupation probabilities and the factorial moments of the incremental loads. We begin with the occupation probabilities, and then turn to the factorial moments.

In terms of the PGF $G\left(z;k,m\right)$ the steady state probability of finding exactly l particles ($l=0,1,2,\cdots$) in the interval $\{k,k+1,\cdots,k+m-1\}$ is given by

$$P_l\left(k,m\right) = \frac{1}{l!}\left.\frac{d^l}{dz^l}G\left(z;k,m\right)\right|_{z=0}\ , \tag{255}$$

with $P_0\left(k,m\right) = G\left(0;k,m\right)$. Taking the l^{th} derivative of Eq. (253) with respect to the variable z, setting $z=0$ and dividing by $l!$, Eq. (255) yields the following recursion for the occupation probabilities

$$\begin{aligned} \mathrm{P}_l(k,m) &= \Pi\left(k,m\right)\delta_{l,0} \\ &+\Pi\left(k,m\right)\sum_{i=1}^{m}\frac{\mu_{k-1}}{\mu_{k-1}+\mu_{k+i-1}}\frac{\mathrm{P}_l(k-1,i+1)}{\Pi(k,i)}\ . \end{aligned} \tag{256}$$

Equation (256), together with the boundary condition in Eq. (248), establishes an explicit iterative scheme for the computation of the occupation probabilities $\mathrm{P}_l(k,m)$ in terms of the occupation probabilities $\{\mathrm{P}_l(k-1,\mathrm{i})\}_{i=2,\cdots,m+1}$.

Analogously, one can further derive recursive equations for the factorial moments of the incremental load $L\left(k,m\right)$. In terms of the PGF $G\left(z;k,m\right)$, the factorial moments $M_l\left(k,m\right)$ ($l=1,2,3,\cdots$) are given by

$$M_l\left(k,m\right) \equiv \left\langle\prod_{i=0}^{l-1}\left(L\left(k,m\right)-i\right)\right\rangle = \left.\frac{d^l}{dz^l}G\left(z;k,m\right)\right|_{z=1}\ . \tag{257}$$

Hence, taking the l^{th} derivative of Eq. (253) with respect to the variable z and setting $z=1$, Eq. (257) yields the following recursive equation for the factorial moments

$$\begin{aligned} M_l\left(k,m\right) &= \\ \Pi\left(k,m\right)&\sum_{i=1}^{m}\frac{\mu_{k-1}}{\mu_{k-1}+\mu_{k+i-1}}\frac{M_l(k-1,i+1)}{\Pi(k,i)}\ . \end{aligned} \tag{258}$$

Equation (258), together with the boundary condition

$$M_l\left(1,m\right)=\left.\frac{d^l}{dz^l}\prod_{i=1}^{m}\frac{1}{1+\frac{\lambda}{\mu_i}\left(1-z\right)}\right|_{z=1} \tag{259}$$

establishes an explicit iterative scheme for the computation of the factorial moments $M_l\left(k,m\right)$ in terms of the factorial moments $\{M_l\left(k-1,i\right)\}_{i=2,\cdots,m+1}$.

8.6 Incremental Load: Exact Results

In this section we return to the analysis of homogeneous ASIPs and provide exact results for the occupation probabilities, factorial moments, and PGF of the incremental load $L(k,m)$. In what follows we return to the convention by which $\mu=1$ and λ is measured in units of the gate opening rate. The results presented in Subsection 8.6.1 are expressed in terms of the entries of Catalan's trapezoids $C_m(n,k)$ (see Chapter 7). Detailed proofs and derivations are given in Subsection 8.6.2.

8.6.1 Occupation Probabilities and Factorial Moments

We start with the incremental load $L\left(1,m\right)$. Substituting $p_i\to 1/(1+\lambda)$ in Eq. (248) we obtain the probabilities $P_l(1,m)$ given by Eq. (222). Similarly, substituting $\lambda/\mu_i\to\lambda$ in Eq. (259) we obtain the corresponding factorial moments

$$M_l\left(1,m\right)=\frac{(m+l-1)!}{(m-1)!}\lambda^l\,. \tag{260}$$

We now turn to the incremental load $L(k,m)$, with $k>1$. In what follows we show that the occupation probabilities $P_l\left(k,m\right)$ $(l=0,1,2,\cdots)$ are given by

$$\begin{aligned}\mathrm{P}_l(k,m)=\delta_{l,0}\sum_{j=2}^{k}\tfrac{C_1(k+m-j-1,k-j)}{2^{2k+m-2j}}\\ +\sum_{j=2}^{k+m-1}\tfrac{C_m(k-2,m+k-1-j)P_l(1,j)}{2^{2k+m-2-j}}\,.\end{aligned} \tag{261}$$

and that the factorial moments $M_l\left(k,m\right)$ $(l=1,2,\cdots)$ are given by

$$M_l\left(k,m\right)=\sum_{j=2}^{k+m-1}\tfrac{C_m(k-2,m+k-1-j)M_l(1,j)}{2^{2k+m-2-j}}\,. \tag{262}$$

We note that the sums in Eqs. (261–262) contain a finite number of explicitly known summands and can thus be used for exact and efficient calculation of $P_l\left(k,m\right)$ and $M_l\left(k,m\right)$. Moreover, in the case of single-site lattice intervals $(m=1)$ the sums in Eqs. (261–262) can be computed (given Eqs. (222) and (260), and by use of any standard computer algebra software) to be expressed

in terms of standard functions. Specifically, the probability distribution and the factorial moments of the random variable X_k are given by

$$\begin{aligned} \mathrm{P}_l(k,1) &= \delta_{l,0}\left(1-\tfrac{\Gamma(k-1/2)}{\sqrt{\pi}\Gamma(k)}\right) \\ &+\tfrac{(1+l)\Gamma(k-3/2)\lambda^l}{2\sqrt{\pi}\Gamma(k)(1+\lambda)^{2+l}} \times {}_2F_1\left(2-k,2+l,4-2k;\tfrac{2}{1+\lambda}\right), \end{aligned} \tag{263}$$

and

$$M_l(k,1) = \tfrac{2^l\lambda^l\Gamma(1+l/2)\Gamma(k+l/2-1/2)}{\sqrt{\pi}\Gamma(k)}, \tag{264}$$

where $\Gamma(x)$ and ${}_2F_1(a,b,c;x)$ are the Gamma function and hypergeometric function, respectively. For large k, an asymptotic analysis of the exact expressions (261) and (263), yields the asymptotic results of Section 8.3. The details of this asymptotic analysis are sketched in Appendix 8.8.8.

8.6.2 The probability generating function

In this subsection we derive an expression for the probability generating function $G(z;k,m)$ and prove the validity of Eqs. (261) and (262). Substituting $\lambda/\mu_i \to \lambda$ in Eq. (246) we see that the probability generating function of the incremental load $L(1,m)$ is given by

$$G(z;1,m) = \left(\frac{1}{1+\lambda(1-z)}\right)^m. \tag{265}$$

We now turn to derive an expression for $G(z;k,m)$ in the case of $k>1$. Our derivation is based on an insightful probabilistic interpenetration of the boundary value problem that appears in Eq. (252) and the main idea behind it is illustrated in Figure 26. An alternative derivation which is algebraic in nature is given in Appendix 8.8.9.

The first step in our derivation is to note that Eq. (252) is linear with respect to the PGFs that compose it. It follows that $G(z;k,m)$ can be expressed as a weighted sum over known boundary PGFs of the type $G(z;1,m)$ and $G(z;k,0)$. Iterating Eq. (252) in an attempt to find the contribution of a specific boundary PGF to the unknown PGF $G(z;k,m)$, we consider a path in the first quadrant of the (k,m) plane that: (i) is composed out of steps in the south ($\downarrow$) and northwest ($\nwarrow$) directions only; (ii) connects the point (k,m) with a specific boundary point (k',m') whose position is associated with the last two arguments of the boundary PGF whose contribution we are trying to assess; (iii) does not pass through any other boundary point. A path that complies with the above-mentioned conditions will henceforth be named a *legitimate path*.

The number of northwest steps in a legitimate path is given by $k-k'$, the number of south steps is given by $k-k'+m-m'$ and the total number of steps is given by $2k-2k'+m-m'$. Since we are dealing with a homogeneous ASIP,

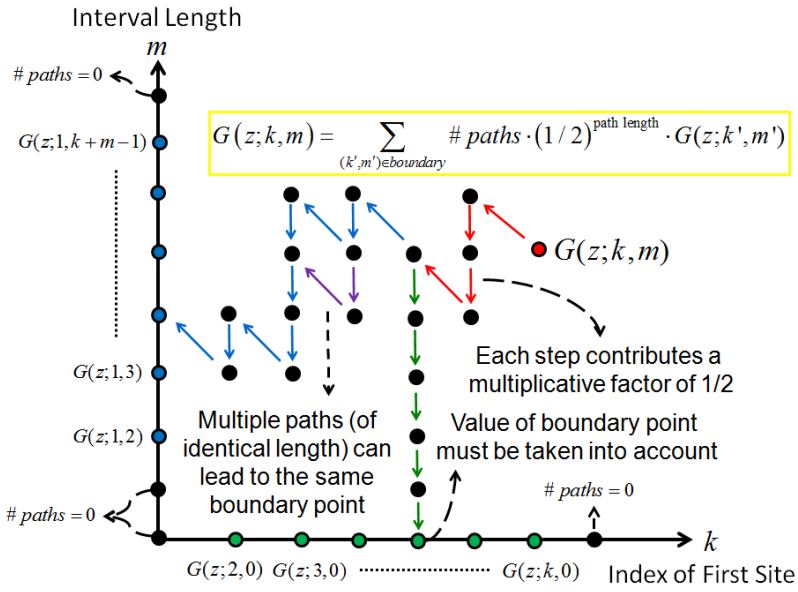

Figure 26: Expressing $G(z;k,m)$ as a weighted sum over known boundary functions. All boundary functions must be properly weighted and taken into account. The weight of each boundary function is given by the number of legitimate paths leading to it, multiplied by 1/2 raised to the power of the path length. Paths are made out of south (↓) and northwest (↖) steps only and must not pass through another boundary function except the one lying at the end of the path. Some boundary functions can be reached via several different paths while others cannot be reached at all (the latter are discarded in the computation of the sum).

Eq. (252) asserts that each step in the path contributes a multiplicative factor of exactly 1/2. The contribution due to a single legitimate path connecting the points (k,m) and (k',m') is hence $(1/2)^{2k-2k'+m-m'}\,G(z;k',m')$. Taking into account all possible legitimate paths and summing over all boundary points we have

$$G(z;k,m) = \sum_{(k',m')\in boundary} \left[\#(k,m,k',m')\right. \\ \left.(1/2)^{2k-2k'+m-m'}\,G(z;k',m')\right], \tag{266}$$

where $\#(k,m,k',m')$ is the number of legitimate paths that start at (k,m) and end at (k',m').

In order to proceed we consider a random walker that chooses, with equal probability at each step, between a south (↓) and northwest step (↖). Assume that the random walker starts its walk at the point (k,m) and let $P_{\rm hit}^{k,m}(k',m')$ be the probability that the random walker hits the boundary point (k',m')

before it hits any other boundary point. From this definition it readily follows that

$$P_{\rm hit}^{k,m}(k',m') = \#(k,m,k',m') \cdot (1/2)^{2k-2k'+m-m'} \ . \tag{267}$$

We will now show that

$$P_{\rm hit}^{k,m}(j,0) = \left(\frac{1}{2}\right)^{2k+m-2j} C_1(k+m-j-1,k-j) \tag{268}$$

$(m = 1,2,\cdots; k = 2,3,\cdots; j = 2,3,\cdots,k)$, and that

$$P_{\rm hit}^{k,m}(1,j) = \left(\frac{1}{2}\right)^{2k+m-2-j} C_m(k-2,m+k-1-j) \tag{269}$$

$(m = 1,2,\cdots; k = 2,3,\cdots; j = 2,3,\cdots,k+m-1)$.

In every legitimate path connecting the point (k,m) with the point $(j,0)$ $(j = 2,3,\cdots,k)$ the last step is always directed to the south. The remaining steps — $k-j$ northwest and $k-j+m-1$ south — must be ordered into a path that connects the point (k,m) to the point $(j,1)$ without hitting the south boundary first. Similarly, in every legitimate path connecting the point (k,m) with the point $(1,j)$ $(j = 2,3,\cdots,k+m-1)$ the last step is always directed to the northwest. The remaining steps — $k-2$ northwest and $k-1+m-j$ south — must be ordered into a path that connects the point (k,m) to the point $(2,j-1)$ without hitting the south boundary first. Recalling the combinatorial interpretation of $C_m(n,k)$ one can easily convince himself that $\#(k,m,j,0) = C_m(k-j,k-j+m-1)$ and $\#(k,m,1,j) = C_m(k-2,k-1+m-j)$. Equation (269) now follows immediately from Eq. (267). Equation (268) follows from Eq. (267) by use of the "diagonal identity" $C_m(k-j,k+m-j-1) = C_1(k+m-j-1,k-j)$. That is, the main diagonal of Catalan's trapezoid of order m coincides with the m^{th} diagonal of Catalan's triangle. This identity is easily verified by use of Eqs. (199) and (202).

The PGF of the incremental load $L(k,m)$ can now be obtained by substituting Eq. (267) into Eq. (266), omitting terms for which $\#(k,m,k',m') = 0$, and utilizing Eqs. (268)-(269) to get

$$\begin{aligned} G\,(z;k,m) &= \sum_{j=2}^{k} G\,(z;j,0) \cdot P_{\rm hit}^{k,m}(j,0) \\ &+ \textstyle\sum_{j=2}^{k+m-1} G\,(z;1,j) \cdot P_{\rm hit}^{k,m}(1,j)\,. \end{aligned} \tag{270}$$

Taking the l^{th} derivative of Eq. (270) with respect to the variable z and setting $z = 1$, Eq. (262) follows by use of Eq. (257) and the fact that $G\,(z;j,0) = P_0(j,0) = 1$. Substituting Eqs. (268)-(269) into Eq. (270) we conclude that

$$\begin{aligned} G\,(z;k,m) &= \sum_{j=2}^{k} G\,(z;j,0) \left(\tfrac{1}{2}\right)^{2k+m-2j} C_1(k+m-j-1,k-j) \\ &+ \textstyle\sum_{j=2}^{k+m-1} G\,(z;1,j) \left(\tfrac{1}{2}\right)^{2k+m-2-j} C_m(k-2,m+k-1-j)\,. \end{aligned} \tag{271}$$

The occupation probabilities in Eq. (261) can then be read off from Eq. (271) after substituting $G(z;j,0)=1$ and $G(z;1,j)=\sum_{l=0}^{\infty}P_l(1,j)z^l$.

8.7 Conclusions and Outlook

In this chapter we studied incremental load probabilities in the ASIP model, analyzed their asymptotic behavior and discussed their implications. Introducing the notion of incremental load, and analyzing it via two complementary approaches — a continuum diffusion-limit approach, and an exact probabilistic-combinatorial approach — we analytically derived expressions for the occupation probabilities of the ASIP's lattice intervals, their corresponding factorial moments, and for the probability distribution of the ASIP's inter-exit time. Spanning both exact results and asymptotic behaviors, the analysis presented herein joins the one in previous chapters to provide the most comprehensive description of the ASIP's steady-state statistics to date.

Looking into the future, this thesis can be viewed as part of a long term goal — the elucidation of the ASIP's steady state distribution in full detail. As an intermediate step, it is natural to turn to the study of correlations between the occupations of several disjoint intervals. The empty interval method was employed to the study of correlations for ASIPs on a ring [55], and may thus also prove useful for open boundary ASIPs. This question is especially interesting in light of the picture discussed above of an open ASIP as a 'conveyor belt': if a single snapshot of an open boundary ASIP is similar to the temporal evolution of the ASIP on a ring, it would be interesting to examine the relation between two-point correlation functions in the former and two-time correlation functions in the latter.

Other interesting questions remain open, many of which are related to the concept of universality. To this end, it would be very interesting to examine the robustness, and inevitable collapse, of our results with respect to a large range of perturbations. For example, it would be interesting to further consider the effect of non-homogeneous hopping rates on cluster formation and delineate the conditions under which non-homogeneity is asymptotically averaged out. While some progress in this direction has already been made in Chapter 6, much remains to be done. Modifying the ASIP a bit, one may ask how does a dependence of the hopping rate on cluster size affect the observed statistics? Another question is what happens when particles arrive to sites other than the first? Finally, the analysis of a generalized ASIP in which hopping times are non-exponential would be both interesting and undoubtedly extremely challenging as it will inevitably require different methods than the ones applied herein.

Before finishing, we wish to take a step back and reexamine the big picture. The ASEP, TJN (asymmetric ZRP) and ASIP are distinct models for unidirectional transport and each of them has already attracted a considerable amount of interest in its own right. On the other hand, the three models are tightly linked as the gate/site capacity parametrization introduced in Chpater 3 revealed that they are all special instances of the same generalized model and

are furthermore unique limiting cases of it. Joined together, they portray a panoramic view which spans the broad spectrum of extremities displayed by unidirectional transport.

Detailed analysis of limiting behavior has proven instrumental in shaping our understanding and "feel" for many physical systems. Indeed, much can be learned by comparing and contrasting quantitative, and perhaps even more importantly qualitative, behavioral features under various limiting regimes. It is for this very reason that a comparative study, aimed at providing a bird's-eye view on unidirectional transport, is timely now more than ever. While in depth discussion along these lines is clearly beyond the scope of this manuscript, some words are still in place. As mentioned above, when exclusion and coagulation are both absent (viz. TJN), site occupancies are statistically independent from one another and the joint probability distribution is characterized by a product form. Moreover, marginal probability distributions are insensitive to the relative position of a site along the lattice and depend only on the inflow and gate opening rates [7, 8, 11, 12]. Interestingly, both exclusion and coagulation ruin these nice properties by inducing spatial correlations and positional dependencies. The final outcome is however markedly different.

In the ASEP, even mean occupancies are sensitive to the location of a site along the lattice [23] but the occupancy profile itself can be captured by a mean field approximation that completely neglects spatial correlations. Situation is different in the ASIP where mean occupancies do not reveal any positional dependencies — even when calculated exactly [1]. Transiently high occupancies "magically" average out with low, keeping the mean occupancy flat (in homogenous ASIPs) and the underlying turmoil hidden. Huge, position dependent, fluctuations are a hallmark of the ASIP and are not seen in the ASEP or TJN. Their emergence can only be understood in light of spatial correlations and their existence dooms any mean-based description profoundly incomplete. So different in nature, one can only wonder how is it possible for the ASEP, TJN and ASIP to be complementary faces of the exact same phenomenon. Fragmented for way too long, unidirectional transport awaits unification.

In the following chapter, we will digress from the main theme of this thesis and discuss the computational modeling of gene translation — a biological process which provides a naturally occurring example for a TSS.

8.8 Appendix

8.8.1 Derivation of Eq. (218)

In this Appendix we present the derivation of Eq. (218). To do so, we define two auxiliary probability functions

$$\begin{aligned}\hat{P}_l^{\rm left}(t;k,m) &\equiv \Pr\Big(L(t;k,m)=l \text{ and } X_{k-1}(t)=0\Big)\\ \hat{P}_l^{\rm right}(t;k,m) &\equiv \Pr\Big(L(t;k,m)=l \text{ and } X_{k+m}(t)=0\Big).\end{aligned} \tag{272}$$

These are the probabilities that sites $\{k,\cdots,k+m-1\}$ are occupied by l particles, and the site immediately to their left/right is empty. Next, note that the probability that sites $\{k-1,\cdots,k+m-1\}$ support l particles and site $k-1$ is *not* empty is exactly $P_l(t;k-1,m+1)-\hat{P}_l^{\rm left}(t;k,m)$. Similarly, the probability that sites $\{k,\cdots,k+m-2\}$ support l particles and site $k+m-1$ is not empty is exactly $P_l(t;k,m-1)-\hat{P}_l^{\rm right}(t;k,m-1)$. Although the auxiliary probabilities are needed in order to write down the equation of motion for $P_l(t;k,m)$,

$$\begin{aligned}\tfrac{\partial}{\partial t}P_l(t;k,m) &= \Big(P_l(t;k-1,m+1)-\hat{P}_l^{\rm left}(t;k,m)\Big)\\ &+\Big(P_l(t;k,m-1)-\hat{P}_l^{\rm right}(t;k,m-1)\Big)\\ &-\Big(P_l(t;k,m)-\hat{P}_l^{\rm left}(t;k,m)\Big)\\ &-\Big(P_l(t;k,m)-\hat{P}_l^{\rm right}(t;k,m-1)\Big),\end{aligned} \tag{273}$$

they cancel out in Eq. (273) and Eq. (218) readily follows.

8.8.2 Derivation of Eq. (220)

The derivation of Eq. (220) is similar to the derivation of Eq. (218) albeit replacing terms corresponding to the entry of particles into the interval from the left (first and third lines of the right hand side of Eq. (273)) with terms corresponding to the arrival of a particle to the first site. The resulting equation is

$$\begin{aligned}\tfrac{\partial}{\partial t}P_l(t;1,m) &= \lambda P_{l-1}(t;1,m)-\lambda P_l(t;1,m)\\ &+\Big(P_l(t;1,m-1)-\hat{P}_l^{\rm right}(t;1,m-1)\Big)\\ &-\Big(P_l(t;1,m)-\hat{P}_l^{\rm right}(t;1,m-1)\Big).\end{aligned} \tag{274}$$

Once again, the auxiliary probabilities cancel out, yielding Eq. (220).

8.8.3 Universality of Eqs. (231) and (232): an explicit example

In this Appendix we demonstrate how the asymptotic scaling forms (231) and (232) emerge for an explicit example of an ASIP with a generalized arrival process. As explained in Section 8.3.4, the universality is a result of the central limit theorem for the distribution of $L(1,k)$, which leads to Eq. (230). The scaling forms are obtained by showing, for the specific example considered below, that the central limit theorem applies. We also present a formal argument that heuristically explains why the central limit theorem is expected to apply for a much larger class of arrival processes.

Consider an ASIP in which particles may enter the first site not only one by one, but also in batches of $n = 1, 2, 3, 4, \ldots$ particles. The arrival of a batch of n particles is assumed to be a Poisson process with rate λ_n. The occupation of the first site thus increases according to the rule

$$X_1, X_2, \cdots \xrightarrow{\lambda_n} X_1 + n, X_2, \cdots , \tag{275}$$

and otherwise the ASIP dynamics remains unchanged. The goal of the current calculation is to find the initial condition $P_l(1,m)$ that is generated by this arrival process, and to analyze the conditions under which the central limit theorem leads to the approximation (230).

The equation equivalent to (220) for this generalized ASIP is

$$\begin{aligned} \tfrac{\partial}{\partial t} P_l(t;1,m) &= \Big[P_l(t;1,m-1) - P_l(t;1,m)\Big] \\ &- \textstyle\sum_{n=1}^{\infty} \lambda_n \Big[P_l(t;1,m) - P_{l-n}(t;1,m)\Big]. \end{aligned} \tag{276}$$

Multiplying by z^l and summing over l leads, in the steady state, to

$$\begin{aligned} &G(z;1,m-1) - G(z;1,m) = \\ &[f_\lambda(1) - f_\lambda(z)] G(z;1,m), \end{aligned} \tag{277}$$

where $f_\lambda(z)$ is the generating function for λ_n:

$$f_\lambda(z) \equiv \sum_{n=1}^{\infty} \lambda_n z^n. \tag{278}$$

Iterating (277) and using $G(z;1,0) = 1$ yields

$$G(z;1,m) = [1 + f_\lambda(1) - f_\lambda(z)]^{-m}. \tag{279}$$

One observes that the distribution of $L(1,m)$ has a product form and is equal to the distribution of a sum of i.i.d. random variable whose generating function is

$$g(z) \equiv [1 + f_\lambda(1) - f_\lambda(z)]^{-1}, \tag{280}$$

compare with Eq. (265). The central limit theorem for this sum applies when the mean and variance of these i.i.d. variables is finite, i.e., when $g'(1), g''(1) < \infty$. It is easy to verify that $g'(1) = f'(1)$ and $g''(1) = 2f''(1) + [f'(1)]^2$. Thus, as long as $\sum_n n^2 \lambda_n < \infty$, one obtains for $m \gg 1$

$$P_l(1,m) \sim \delta(l - m\langle\lambda\rangle) = \langle\lambda\rangle^{-1}\delta(m - l/\langle\lambda\rangle), \tag{281}$$

where we have defined $\langle\lambda\rangle \equiv \sum_n n\lambda_n$.

Let us now motivate in a heuristic fashion why the central limit theorem is expected to hold for a much larger class of arrival processes. Assume that the arrival process is such that Eq. (277) is replaced by

$$G(z;1,m-1) - G(z;1,m) = \mathcal{A}(z)G(z;1,m), \tag{282}$$

where $\mathcal{A}(z)$ is a formal notation for the operator associated with the arrival process. The formal solution of this equation is $G(z;1,m) = [1+\mathcal{A}(z)]^{-m}$ [compare with Eq. (279)]. If the operator $[1+\mathcal{A}(z)]^{-1}$ is characterized by a non-vanishing spectral gap, i.e., there is a finite difference between its largest and second-largest eigenvalues, then when $m \to \infty$ one has asymptotically $G(z;1,m) \sim g_{\max}(z)^m$, where $g_{\max}(z)$ denotes the largest eigenvalue of $[1+\mathcal{A}(z)]^{-1}$ for some fixed value of z. If, in addition, $g_{\max}(z)$ is the PGF of a random variable with finite variance, a central limit theorem holds for $L(1,m)$ and an approximation of the form (230) is valid.

8.8.4 Saddle point evaluation of Eq. (239)

In this section we show how Eq. (209) follows by applying a saddle point approximation (also known as Laplace's method) to the sum in Eq. (239) in the limit $k \to \infty$. The first step is to apply Stirling's approximation to the probability density of the traversal time

$$\frac{t^m e^{-t}}{m!} \simeq \frac{e^{-t+m\log(t/m)+m}}{\sqrt{2\pi m}}\,. \tag{283}$$

Next, we substitute Eqs. (283) and (237) into Eq. (239) to obtain

$$P_{T_k}(t) \simeq \sum_{m=1}^{k-1} \frac{e^{-t+m\log(t/m)+m}}{\sqrt{2\pi m}} \frac{m e^{-m^2/4(k-m)}}{2(k-m)}\,. \tag{284}$$

Setting $u = m/\sqrt{k}$ we rewrite (284) as

$$P_{T_k}(t) \simeq \sum_{u} \frac{e^{-t+u\sqrt{k}\log(t/u\sqrt{k})+u\sqrt{k}}}{\sqrt{2\pi u\sqrt{k}}} \frac{u e^{-u^2/4\left(1-u/\sqrt{k}\right)}}{2(\sqrt{k}-u)}\,, \tag{285}$$

where the sum runs over values $u = k^{-1/2}, 2k^{-1/2}, \cdots, k^{1/2} - k^{-1/2}$. We now observe that

$$P_{T_k}(\sqrt{k}t) \simeq \sum_u \frac{k^{1/4}u^{1/2}}{\sqrt{8\pi}(k-\sqrt{k}u)} e^{\sqrt{k}f(u)} \tag{286}$$

with

$$f(u) \equiv u\log(t/u) + u - t - u^2/(4\sqrt{k} - 4u)\,. \tag{287}$$

For large k, the sum in Eq. (286) may be approximated by an integral, which can be evaluated using a saddle point approximation. We thus search for a saddle point u^* for which $f'(u*) = 0$ and find it to be

$$u^* = t - t^2/2\sqrt{k} + O(k^{-1})\,, \tag{288}$$

(u^* is computed to leading order in k such that $\lim_{k\to\infty} \sqrt{k}f'(u^*) = 0$). Evaluating the integral approximation of the sum in Eq. (286) to leading order, we find

$$\begin{aligned} P_{T_k}(\sqrt{k}t) &\simeq \textstyle\int_0^{\sqrt{k}} \frac{k^{3/4}u^{1/2}}{\sqrt{8\pi}(k-\sqrt{k}u)} e^{\sqrt{k}f(u)} du \\ &= \tfrac{te^{-t^2/4}}{2\sqrt{k}} + O(k^{-1})\,. \end{aligned} \tag{289}$$

We now observe that the probability density function of the normalized inter-exit time $T_k/\sqrt{\pi k}$ is related to the probability density function of T_k in the following way

$$P_{T_k/\sqrt{\pi k}}(t) = \sqrt{\pi k} P_{T_k}(\sqrt{\pi k}t)\,. \tag{290}$$

Equation (209) follows immediately.

8.8.5 Derivation of Eq. (244)

Conditioning on the occupancy vector $\mathbf{X}(t)$ and utilizing the Markovian dynamics of Eq. (243) we have

$$\begin{aligned} \left\langle z^{L\left(t';1,m\right)} \right\rangle &= \left\langle \left\langle z^{L\left(t';1,m\right)} | \mathbf{X}(t) \right\rangle \right\rangle \\ &= \begin{cases} (\lambda\Delta) \left\langle z^{L(t;1,m)+1} \right\rangle \\ + \\ (\mu_m\Delta) \left\langle z^{L(t;1,m-1)} \right\rangle \\ + \\ (1 - (\lambda+\mu_m)\,\Delta) \left\langle z^{L(t;1,m)} \right\rangle \\ + \\ o(\Delta)\,. \end{cases} \end{aligned} \tag{291}$$

Equation (244) is obtained after rearranging terms in Eq. (291), dividing by Δ, taking $\Delta \to 0$ and using the PGF notation of Eq. (240).

8.8.6 Derivation of Eq. (251)

Conditioning on the occupancy vector $\mathbf{X}(t)$ and utilizing the Markovian dynamics of Eq. (250) we have

$$\begin{aligned}\left\langle z^{L\left(t';k,m\right)}\right\rangle &= \left\langle \left\langle z^{L\left(t';k,m\right)}|\mathbf{X}(t)\right\rangle \right\rangle \\ &= \begin{cases} \left(\mu_{k-1}\Delta\right)\left\langle z^{L(t;k-1,m+1)}\right\rangle \\ + \\ \left(\mu_{k+m-1}\Delta\right)\left\langle z^{L(t;k,m-1)}\right\rangle \\ + \\ \left(1-\left(\mu_{k-1}+\mu_{k+m-1}\right)\Delta\right)\left\langle z^{L(t;k,l)}\right\rangle \\ + \\ o(\Delta)\ . \end{cases}\end{aligned} \tag{292}$$

Equation (251) is obtained after rearranging terms in Eq. (292), dividing by Δ, taking $\Delta \to 0$ and using the PGF notation of Eq. (240).

8.8.7 Derivation of Eq. (253)

We prove Eq. (253) by showing that the probability generating function $G\left(z;k,m\right)$ it defines satisfies Eq. (252). To this end we apply mathematical induction on the index k. We start by showing that Eq. (253) holds for $k=2$ and an arbitrary value of m. Indeed, for $k=2$ Eq. (253) reads

$$\begin{aligned} G\left(z;2,m\right) &= \Pi\left(2,m\right) \\ &+\Pi\left(2,m\right)\textstyle\sum_{j=1}^{m}\frac{\mu_1}{\mu_1+\mu_{1+j}}\frac{G(z;1,j+1)}{\Pi(2,j)}\ . \end{aligned} \tag{293}$$

Substituting Eq. (293) into Eq. (252) and utilizing Eq. (246) we have

$$\begin{aligned} &\Pi\left(2,m\right)\left(1+\textstyle\sum_{j=1}^{m}\frac{\mu_1}{\mu_1+\mu_{1+j}}\frac{\prod_{i=1}^{j+1}\frac{\mu_i}{\mu_i+\lambda(1-z)}}{\Pi(2,j)}\right)\stackrel{?}{=} \\ &+\tfrac{\mu_{1+m}}{\mu_1+\mu_{1+m}}\Pi\left(2,m-1\right)\left(1+\textstyle\sum_{j=1}^{m-1}\frac{\mu_1}{\mu_1+\mu_{1+j}}\frac{\prod_{i=1}^{j+1}\frac{\mu_i}{\mu_i+\lambda(1-z)}}{\Pi(2,j)}\right) \\ &+\tfrac{\mu_1}{\mu_1+\mu_{1+m}}\textstyle\prod_{j=1}^{m+1}\frac{\mu_j}{\mu_j+\lambda(1-z)}\ . \end{aligned} \tag{294}$$

Canceling matching terms on both sides of Eq. (294) gives the trivial identity $0=0$ and proves our claim.

We finish the proof by showing that if Eq. (253) holds for $k \geq 2$ it holds for $k+1$ as well. Indeed, replacing k by $k+1$ in Eqs. (252-253) we substitute Eq. (253) into Eq. (252) and obtain

$$\Pi\left(k+1,m\right)\left(1+\sum_{j=1}^{m}\frac{\mu_k}{\mu_k+\mu_{k+j}}\frac{G(z;k,j+1)}{\Pi(k+1,j)}\right)\stackrel{?}{=}$$

$$+\frac{\mu_{k+m}\Pi(k+1,m-1)}{\mu_k+\mu_{k+m}}\left(1+\sum_{j=1}^{m-1}\frac{\mu_k}{\mu_k+\mu_{k+j}}\frac{G(z;k,j+1)}{\Pi(k+1,j)}\right) \tag{295}$$

$$+\frac{\mu_k\Pi(k,m+1)}{\mu_k+\mu_{k+m}}\left(1+\sum_{j=1}^{m+1}\frac{\mu_{k-1}}{\mu_{k-1}+\mu_{k+j-1}}\frac{G(z;k-1,j+1)}{\Pi(k,j)}\right).$$

Canceling matching terms on both sides gives

$$G\left(z;k,m+1\right)=$$

$$\Pi\left(k,m+1\right)\left(1+\sum_{j=1}^{m+1}\frac{\mu_{k-1}}{\mu_{k-1}+\mu_{k+j-1}}\frac{G(z;k-1,j+1)}{\Pi(k,j)}\right) \tag{296}$$

which coincides with Eq. (253) for $G\left(z;k,m+1\right)$ and concludes our proof.

8.8.8 Asymptotic analysis of Equation (261)

In this Appendix we sketch the asymptotic analysis of the exact expression for the occupation probabilities (261) and show how the results of Section 8.3, and in particular Eqs. (203), (228)–(229), and (232) can be obtained from it. We concentrate here solely on the case of $m=1$; the calculation for other values of m is similar but somewhat more lengthy.

For the case $m=1$, the sum in (261) can be rewritten, by substituting the definition (202) and the "initial condition" (222), as

$$P_0(k,1)=\sum_{i=0}^{k-2}\binom{2i+1}{i}\frac{1}{2i+1}2^{-(2i+1)}+$$
$$+\sum_{i=0}^{k-2}\binom{k-1+i}{i}\frac{k-1-i}{k-1+i}2^{-(k-1+i)}(1+\lambda)^{-(k-i)} \tag{297}$$

for $l=0$, while for $l\geq 1$ it has the form

$$P_l(k,1)=S(2,k), \tag{298}$$

where we define

$$S(j_1,j_2)\equiv\Big(\frac{\lambda}{1+\lambda}\Big)^l\sum_{j=j_1}^{j_2}\frac{j-1}{2k-j-1}\binom{2k-j-1}{k-1}\times \tag{299}$$
$$\times\binom{l+j-1}{l}2^{-(2k-j-1)}(1+\lambda)^{-j}.$$

To obtain these relations we have used the binomial identity $\binom{n-1}{k}-\binom{n}{k}=\frac{n-2k}{n}\binom{n}{k}$.

We first evaluate the sums in Eq. (297) for large k. The first sum can be calculated exactly, and equals $1-\Gamma(k-1/2)/\sqrt{\pi}\Gamma(k)\simeq 1-1/\sqrt{\pi k}$. The main contribution to the second sum is from values of i which are close to k. It can be shown, by expanding the summand for $k\gg k-i$, that the second sum decays to zero as $k^{-3/2}$ and is therefore negligible compared to the first. We thus arrive at Eq. (203).

We now move on to the asymptotic evaluation of (298)–(299) for large k. The main contribution to the sum, as shown below, is from values of j which are close to l. Therefore, two cases are treated separately: (i) $l\ll\sqrt{k}$, and (ii) $l=x\sqrt{k}$ with $x=O(1)$.

Case (i), $l\ll\sqrt{k}$. In this case, the sum is evaluated using Stirling's approximation

$$2^{-(2k-j-1)}\binom{2k-j-1}{k-1}\simeq\sqrt{\frac{2k-j-1}{2\pi(k-j)(k-1)}}e^{-f_1(k)}, \tag{300}$$

with

$$f_1(k)=(k-1)\log\frac{k-1}{2k-j-1}+(k-j)\log\frac{k-j}{2k-j-1} \tag{301}$$

$$=\frac{j^2}{4k}\Big[1+O\Big(\frac{j}{k}\Big)\Big]. \tag{302}$$

The term f_1 in the exponent yields a significant contribution to the summand only for values of j which are comparable with $\sqrt{k}$, while for $j\ll\sqrt{k}$ it is negligible. Accordingly, we split the sum in (298) into two: $P_l(k,1)=S(2,N)+S(N+1,k)$, the first running over $j=2,\ldots,N$, and the second over $j=N+1,\ldots k$. Here $N=N(k)$ is chosen in such a way that $l\ll N\ll k$. The first of these sums may be approximated using (300) as

$$S(2,N)\simeq\Big(\frac{\lambda}{1+\lambda}\Big)^l\sum_{j=2}^{N}\frac{j-1}{\sqrt{4\pi}k^{3/2}}\binom{l+j-1}{l}(1+\lambda)^{-j}. \tag{303}$$

Since $N\gg 1$ and the summand decays exponentially with j, replacing the upper boundary in the last sum by ∞ results in a negligible error. The sum can now be computed exactly, and yields

$$S(2,N)\simeq\frac{l+1}{\sqrt{4\pi}\lambda^2k^{3/2}}. \tag{304}$$

The contribution of the second sum (from $N+1$ to k) is negligible as long as $l\ll\sqrt{k}$. To see this, approximate $\binom{l+j-1}{l}\simeq j^{l+1}/l!$ (which is valid for $j>N\gg l$), and then approximate the sum as an integral:

$$S(N+1,k)\simeq\simeq\Big(\frac{\lambda}{1+\lambda}\Big)^l\frac{k^{(l-1)/2}}{\sqrt{4\pi}l!}\int_{\frac{N}{\sqrt{k}}}^{\sqrt{k}}y^{l+2}e^{-y^2/4-y\sqrt{k}\log(1+\lambda)}dy, \tag{305}$$

where a change of integration variable $y=j/\sqrt{k}$ was made. Once again, we incur a negligible error by approximating the lower and upper integration boundaries

as $N/\sqrt{k} \simeq 0$ and $\sqrt{k} \simeq \infty$. By evaluating the integral, it can be shown that $S(N+1,k) \ll S(2,N)$, leading to (228)–(229) (remember that here $m=1$).

Case (ii), $l = x\sqrt{k}$. In this case, since $l \gg 1$, one may employ Stirling's approximation also for the second binomial coefficient in (299). Replacing as before the sum by an integral with an integration variable $y = j/\sqrt{k}$ leads to

$$P_l(k,1) \simeq \int_0^\infty \Big(\frac{\lambda}{1+\lambda}\Big)^l \frac{y^{3/2}}{4\pi k^{3/4}\sqrt{x(x+y)}} e^{-\frac{y^2}{4}-\sqrt{k}f_2(y)} \tag{306}$$

with

$$f_2(y) \equiv x\log\frac{x}{x+y} + y\log\frac{y}{x+y} + y\log(1+\lambda). \tag{307}$$

For $\sqrt{k} \gg 1$, the integral can be evaluated using a saddle point approximation: f_2 has a minimum at $y^* = x/\lambda$, where its value is $f_2(y^*) = x\log\lambda/(1+\lambda)$. We therefore obtain the scaling form

$$P_l(k,1) \simeq \frac{x}{\sqrt{4\pi}\lambda^2 k} e^{-x^2/4\lambda^2} \tag{308}$$

[compare with (232)]. Note that this saddle point calculation carries through to any $1 \ll l \ll k$. The results are different, however, at the scale of $l = O(k)$, as the main contribution to the sum (the saddle point) comes from values $j = O(k)$, leading to non-negligible corrections to the calculation due to terms neglected above such as the higher order terms in (302).

8.8.9 Derivation of Eq. (271)

In this Appendix we provide an alternative derivation of Eq. (271). The derivation of this Appendix is algebraic in nature and serves to show that the desired result may also be obtained without reference to the probabilistic argumentation presented in the main text. The proof is divided into three parts. In Part I we show that for $k > 1$, $G(z;k,m)$ can be written as

$$\begin{aligned} G(z;k,m) &= \sum_{j=2}^{k} \left(\tfrac{1}{2}\right)^{2k+m-2j} C_1(k+m-j-1,k-j) \\ &+ \left(\tfrac{1}{2}\right)^{2k+m-2} \sum_{l=0}^{\infty} A(k,m,l) z^l \end{aligned} \tag{309}$$

where

$$\begin{aligned} A(k,m,l) = \Big(\tfrac{\lambda}{1+\lambda}\Big)^l \textstyle\sum_{j_1=1}^{m} \sum_{j_2=1}^{j_1+1} \sum_{j_3=1}^{j_2+1} \cdots \\ \cdots \sum_{j_{k-2}=1}^{j_{k-3}+1} \sum_{j_{k-1}=1}^{j_{k-2}+1} \binom{j_{k-1}+l}{j_{k-1}} \Big(\tfrac{2}{1+\lambda}\Big)^{j_{k-1}+1} . \end{aligned} \tag{310}$$

In Part II we show that

$$\begin{aligned} A(k,m,l) = \\ \textstyle\sum_{j=1}^{m+k-2} 2^{j+1} C_m(k-2, m+k-2-j) P_l(1, j+1) . \end{aligned} \tag{311}$$

In Part III we combine Eqs. (309) and (311) to conclude the proof.

Part I. We prove Eq. (309) by induction on k. We start by showing that Eq. (309) holds for $k = 2$ and an arbitrary value of m. Setting $\mu_i = \mu$ $(i = 1, 2, 3, \cdots)$ in Eq. (253) we have

$$\begin{array}{l} G(z;k,m) = \left(\frac{1}{2}\right)^m \\ + \sum_{j=1}^{m} \left(\frac{1}{2}\right)^{m+1-j} G(z;k-1,j+1) \end{array} \tag{312}$$

$(k > 1)$. Setting $k = 2$ in Eq. (312) and utilizing Eq. (265) we have

$$\begin{array}{l} G(z;2,m) = \left(\frac{1}{2}\right)^m \\ + \sum_{j=1}^{m} \left(\frac{1}{2}\right)^{m+1-j} \left(\frac{1}{1+\lambda(1-z)}\right)^{j+1} . \end{array} \tag{313}$$

Recalling the Taylor expansion

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i \tag{314}$$

$|x| < 1$, we expand the parenthesis in the second term of Eq. (313) to obtain

$$\begin{array}{l} G(z;2,m) = \left(\frac{1}{2}\right)^m \\ + \sum_{j=1}^{m} \left(\frac{1}{2}\right)^{m+1-j} \left(\frac{1}{1+\lambda}\right)^{j+1} \left(\sum_{i=0}^{\infty} \left(\frac{\lambda z}{1+\lambda}\right)^i\right)^{j+1} . \end{array} \tag{315}$$

Noting that

$$\left(\sum_{i=0}^{\infty} \left(\frac{\lambda z}{1+\lambda}\right)^i\right)^{j+1} = \sum_{l=0}^{\infty} \left(\begin{array}{c} j+l \\ j \end{array}\right) \left(\frac{\lambda z}{1+\lambda}\right)^l \tag{316}$$

and

$$A(2,m,l) = \left(\frac{\lambda}{1+\lambda}\right)^l \sum_{j=1}^{m} \left(\frac{2}{1+\lambda}\right)^{j+1} \left(\begin{array}{c} j+l \\ j \end{array}\right) \tag{317}$$

we substitute Eq. (316) into Eq. (315) to obtain

$$G(z;2,m) = \left(\frac{1}{2}\right)^m + \left(\frac{1}{2}\right)^{m+2} \sum_{l=0}^{\infty} A(2,m,l) z^l . \tag{318}$$

Noting that $C_1(m-1,0) = 1$ $(m = 1, 2, 3, \cdots)$, we see that Eq. (318) identifies with Eq. (309) for $k = 2$.

We finish the first part of the proof by showing that if Eq. (309) holds for $k \geq 2$ it holds for $k+1$ as well. Indeed, replacing k by $k+1$ in Eq. (312) we substitute Eq. (309) into Eq. (312) and obtain

$$\begin{aligned} & G\left(z; k+1, m\right) = \\ & + \left(\tfrac{1}{2}\right)^m \left[1 + \sum_{i=1}^{m} \sum_{j=2}^{k} \left(\tfrac{1}{2}\right)^{2k+2-2j} C_1(k-j+i, k-j)\right] \\ & + \left(\tfrac{1}{2}\right)^{m+2k} \sum_{i=1}^{m} \sum_{l=0}^{\infty} A(k, i+1, l) z^l \ . \end{aligned} \tag{319}$$

Performing an index shift $j \to k+1-j$, Eq. (319) can be rewritten as

$$\begin{aligned} & G\left(z; k+1, m\right) = \\ & + \left(\tfrac{1}{2}\right)^m \left[1 + \sum_{i=1}^{m} \sum_{j=1}^{k-1} \left(\tfrac{1}{2}\right)^{2j} C_1(i+j-1, j-1)\right] \\ & + \left(\tfrac{1}{2}\right)^{m+2k} \sum_{i=1}^{m} \sum_{l=0}^{\infty} A(k, i+1, l) z^l \ . \end{aligned} \tag{320}$$

We now note that Eqs. (201) and (310) imply respectively that

$$C_1(j+m-1, j) = \sum_{i=1}^{m} C_1(i+j-1, j-1) \tag{321}$$

and

$$A(k+1, m, l) = \sum_{i=1}^{m} A(k, i+1, l) \, . \tag{322}$$

Substituting Eqs. (321) and (322) into Eq. (320) we obtain

$$\begin{aligned} & G\left(z; k+1, m\right) = \\ & + \left(\tfrac{1}{2}\right)^m \left[1 + \sum_{j=1}^{k-1} \left(\tfrac{1}{2}\right)^{2j} C_1(j+m-1, j)\right] \\ & + \left(\tfrac{1}{2}\right)^{m+2k} \sum_{l=0}^{\infty} A(k+1, m, l) z^l \ . \end{aligned} \tag{323}$$

Applying the index shift $j \to k-j+1$ and noting again that $C_1(m-1, 0) = 1$

$(m = 1, 2, 3, \cdots)$ we conclude that

$$\begin{aligned} & G\,(z; k+1, m) = \\ & + \sum_{j=2}^{k+1} \left(\tfrac{1}{2}\right)^{2k+m+2-2j} C_1(k+m-j, k+1-j) \\ & + \left(\tfrac{1}{2}\right)^{m+2k} \sum_{l=0}^{\infty} A(k+1, m, l) z^l \ , \end{aligned} \tag{324}$$

a form which coincides with Eq. (309) for $G\,(z; k+1, m)$.

Part II. We will now prove Eq. (311). Examining Eq. (310) it is easy to see that it can be rewritten in the following form

$$A(k, m, l) = \textstyle\sum_{j=1}^{m+k-2} 2^{j+1} \#_{k,j}^{m} P_l(1, j+1)\,, \tag{325}$$

where we have used Eq. (222) and defined

$$\begin{aligned} \#_{k,j}^{m} = & \textstyle\sum_{j_1=1}^{m} \sum_{j_2=1}^{j_1+1} \sum_{j_3=1}^{j_2+1} \cdots \\ & \cdots \textstyle\sum_{j_{k-2}=1}^{j_{k-3}+1} \sum_{j_{k-1}=1}^{j_{k-2}+1} \delta(j_{k-1}, j) \end{aligned} \tag{326}$$

to be the exact number of times that the running index j_{k-1} in Eq. (310) is equal to j $(j = 1, \cdots, m+k-2)$.

What can be said about the numbers $\#_{k,j}^{m}$? First, it is fairly straightforward to see that when $k = 2$ we have

$$\#_{2,j}^{m} = 1 \tag{327}$$

$(m = 1, 2, \cdots; j = 1, \cdots, m)$. In addition when $j = m+k-2$ we have

$$\#_{k,m+k-2}^{m} = 1 \tag{328}$$

$(m = 1, 2, \cdots; k = 2, 3, \cdots)$. Now, for $k > 2$ and $1 \le j < m+k-2$ we note that the following recursion relation holds

$$\#_{k,j}^{m} = \#_{k-1,j-1}^{m} + \#_{k,j+1}^{m}\,. \tag{329}$$

Indeed, substituting Eq. (326) into Eq. (329) we have

$$\begin{aligned} \#_{k,j}^{m} \stackrel{?}{=} & \textstyle\sum_{j_1=1}^{m} \sum_{j_2=1}^{j_1+1} \sum_{j_3=1}^{j_2+1} \cdots \sum_{j_{k-2}=1}^{j_{k-3}+1} \delta(j_{k-2}, j-1) \\ & + \textstyle\sum_{j_1=1}^{m} \sum_{j_2=1}^{j_1+1} \sum_{j_3=1}^{j_2+1} \cdots \sum_{j_{k-2}=1}^{j_{k-3}+1} \sum_{j_{k-1}=1}^{j_{k-2}+1} \delta(j_{k-1}, j+1) \end{aligned} \tag{330}$$

which immediately gives

$$\begin{aligned}\#^{m}_{k,j} &\stackrel{?}{=} \textstyle\sum_{j_1=1}^{m}\sum_{j_2=1}^{j_1+1}\sum_{j_3=1}^{j_2+1}\cdots \\ &\cdots\textstyle\sum_{j_{k-2}=1}^{j_{k-3}+1}\Big[\delta(j_{k-2},j-1)+\sum_{j_{k-1}=1}^{j_{k-2}+1}\delta(j_{k-1},j+1)\Big]\end{aligned} \tag{331}$$

However, it is easy to check that

$$\sum_{j_{k-1}=1}^{j_{k-2}+1}\delta(j_{k-1},j+1)=\left(\sum_{j_{k-1}=1}^{j_{k-2}+1}\delta(j_{k-1},j)\right)-\delta(j_{k-2},j-1) \tag{332}$$

substituting Eq. (332) into Eq. (331) we recover Eq. (326) and assert the validity of Eq. (329).

We now note that

$$\#^{m}_{k,j}=C_m(k-2,m+k-2-j)\,. \tag{333}$$

Indeed, for $k=2$

$$\#^{m}_{2,j}=C_m(0,m-j)=1 \tag{334}$$

$(m=1,2,\cdots;j=1,\cdots,m)$. In addition, for $j=m+k-2$ we have

$$\#^{m}_{k,m+k-2}=C_m(k-2,0)=1 \tag{335}$$

$(m=1,2,\cdots;k=2,3,\cdots)$. Finally we note that, for $k>2$ and $1\le j<m+k-2$, Eqs. (329) and (333) imply that

$$\begin{aligned}&C_m(k-2,m+k-2-j)=C_m(k-3,m+k-2-j)\\ &+C_m(k-2,m+k-3-j)\end{aligned} \tag{336}$$

which together with the boundary conditions specified in Eqs. (334–335) give back the iterative construction of the Catalan trapezoid of order m. Substituting Eq. (333) into Eq. (325) we recover Eq. (311) and conclude the second part our proof.

Part III. In this part we complete the derivation of Eq. (271). Substituting Eq. (311) into Eq. (309) we have

$$\begin{aligned}G\,(z;k,m)&=\sum_{j=2}^{k}\left(\tfrac{1}{2}\right)^{2k+m-2j}C_1(k+m-j-1,k-j)\\ &+\sum_{l=0}^{\infty}\left[\sum_{j=1}^{m+k-2}P_l(1,j+1)\cdot\left(\tfrac{1}{2}\right)^{2k+m-3-j}C_m(k-2,m+k-2-j)\right]z^l\end{aligned} \tag{337}$$

where we have utilized the fact that $P_0(j,0)=1$. Shifting the index of summation in the inner sum of the second line of Eq. (337) we obtain Eq. (271).

9 Genome-Scale Analysis of Translation Elongation Based on a Ribosome Flow Model

Gene translation is a complex process through which an mRNA sequence is decoded by the ribosome to produce a specific protein. The elongation step of this process is an iterative procedure in which each codon in the mRNA sequence is recognized by a specific tRNA, which adds one additional amino-acid to the growing peptide [84]. As gene translation is a central process in all living organisms, its understanding has ramifications to human health [85, 86, 87], biotechnology [88, 89, 90, 91, 92, 93, 94, 95] and evolution [87, 90, 94, 96].

In recent years there has been a sharp growth in the number of new technologies for measuring different features related to the process of gene translation [88, 89, 93, 97, 98, 99, 100, 101, 102]. However, this process is still enigmatic, with contradicting conclusions in different studies. In particular, the identity of the essential parameters that determine translation rates is still under debate [89, 103, 104]. Recent studies have suggested that the order of codons along the mRNA (and not only the composition of codons) plays an important role in determining translation efficiency [90, 103, 105, 106]. Starting with the seminal work of MacDonald et al. [24] and the work of Heinrich et al. [107] theoretical models for the movement of ribosomes (and other biological 'machines') have been presented [108, 109, 110]. Despite being relatively realistic these models haven't been used for the analysis of large scale genomic data. The models that have been used for this purpose, while making promising and worthy first strides, have not attempted to capture the nature of the translation elongation process on all its various physical aspects [89, 96, 108, 109, 110, 111, 112].

The most widely used predictors of translation efficiency are the codon adaptation index (CAI) [110] and the tRNA adaptation index (tAI) [109]. As we describe later, the tAI is the mean adaptation of a gene (i.e., of its codons) to the tRNA pool of the organism. The CAI is similar to the tAI albeit in this predictor the weight of each codon is computed based on its frequency in a set of highly expressed genes. Based on measures such as the tAI, it is possible to estimate the translation rate of single codons. Thus, it possible to study (local) translation rate profiles along genes [90, 113]. As we depict later, in this chapter we take into account some additional physical aspects of translation elongation.

The aim of the present research is twofold. First, we address the need for a simple, physically plausible computational model that is solely based on the coding sequence (i.e. a vector of codons in each gene). In addition we further require that the model will allow for a computationally efficient analysis of the translation process on a genome-wide scale and across many species. Focusing on the coding sequence, we by no means wish to imply that it is the only factor taking place in the determination of translation rates. Nevertheless, since it has been widely recognized as a prime factor in the translation elongation process, we will herby study it in isolation. To this end, we introduce a new approach for modeling translation elongation. Our model is aimed at capturing the effect of codon order on translation rates, the stochastic nature of the translation

process and the interactions between ribosomes. We demonstrate that our approach gives more accurate predictions of translation rates, protein abundances and ribosome densities in endogenous and heterologous genes in comparison to contemporary approaches.

Second, using our model, we address the need for a better understanding of the translation process. Our analysis unravels several central and yet uncharacterized aspects of this process.

9.1 The Ribosome Flow Model

Our model is based on the Asymmetric Simple Exclusion Process (ASEP, see, for example, [24, 107] and subsequent studies [30]. In the ASEP, initiation times, as well as the time a ribosome spends translating each codon, are exponentially distributed (mean translation times are of course codon dependent). In addition, ribosomes span over several codons and if two ribosomes are adjacent, the trailing one is delayed until the ribosome in front of it has proceeded onwards (see Figure 27A, Subsection 9.5.1, and Subsection 9.5.18).

Despite its rather simple description, the mathematical tractability of the model described above is poor and full, large scale, simulations of it are relatively slow. In order to allow for analytical treatment and in order to reduce simulation times, we introduced two simplifications. First, instead of describing the dynamics at the level of a single mRNA molecule we describe the dynamics after it was averaged over many identical mRNA molecules (see Subsection 9.5.2). Second, we limit ourselves to a spatial resolution that is of the size of a single ribosome. These simplifications will be further explained and justified later.

The simplified model, entitled Ribosome Flow Model (RFM), is illustrated in Figure 27B-C. mRNA molecules are coarse-grained into sites of C codons each; (in Figure 27B $C = 3$); in practice, as we discuss with more details latter, we use $C = 25$ (unless otherwise mentioned), a value that is close to various geometrical properties of the ribosome such as its footprint on the mRNA sequence and the length of its exit channel [90, 97, 105, 114, 115, 116]. As we report later, the choice $C = 25$ is not arbitrary and was made since it gives the best predictions of protein abundance levels.

Ribosomes arrive at the first site with initiation rate λ, but are only able to bind if this site is not occupied by another ribosome. The initiation rate is a function of physical features such as the number of available free ribosomes [90, 117, 118], the folding energy of the 5'UTRs [89, 103], the folding energy at the beginning of the coding sequence [89, 103, 119, 120] and the base pairing potential between the 5'UTR and the ribosomal rRNA [121]. As some of these features and their combined effect are unknown and out of the scope of this paper, we assume a global initiation rate or infer the initiation rate from the coding sequences (as we show in the Subsection 9.3.4). We do so for the sake of simplicity and in order to avoid over-fitting of data.

A ribosome that occupies the $i-th$ site moves, with rate λ_i, to the consecutive site provided the latter is not occupied by another ribosome. Transition

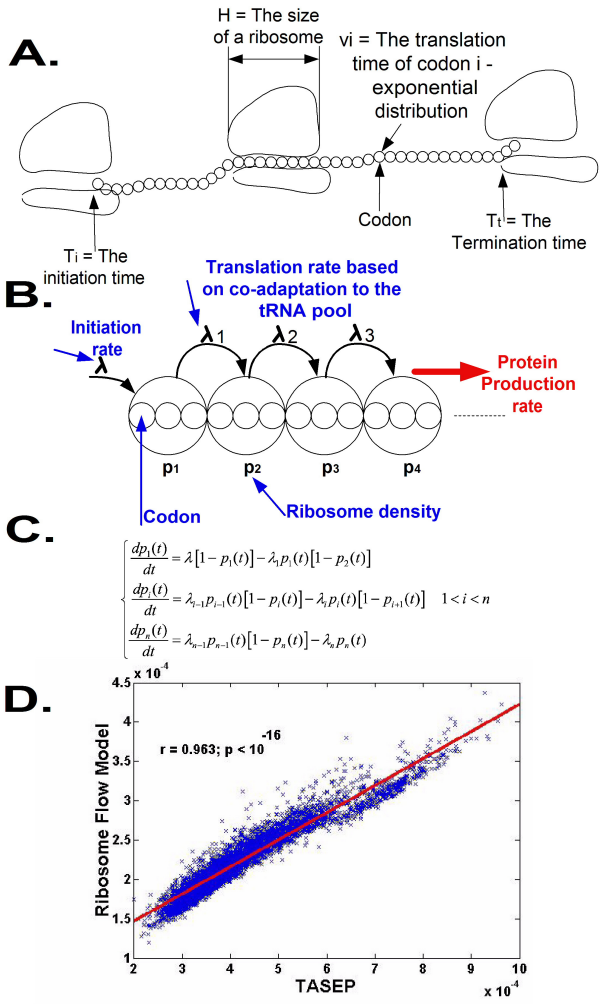

Figure 27: Basic properties of the Ribosome Flow Model (RFM). A. The ASEP model: each codon has an exponentially distributed translation time; ribosomes have volume and can block each other. B. The RFM has two free parameters: the initiation rate λ and the number of codons C at each 'site' (proportional to the size of the ribosome). Each site has a corresponding transition rate that is estimated based on the co-adaptation between the codons of the site and the tRNA pool of the organism. The output of the model consists of the steady state occupancy probabilities of ribosomes at each site and the steady state translation rates, or ribosome flow through the system. C. The set of differential equations that describe the RFM, denoted as Eq. (1). D. RFM vs. ASEP: the correlation between translation rates predicted by the two models is close to perfect ($r = 0.963$, $p < 10 - 16$) while the running time of the ASEP is orders of magnitude longer (usually several days vs. minutes).

rates are determined by the codon composition of each site and the tRNA pool of the organism. Briefly, taking into account the affinity between tRNA species and codons, the translation rate of a codon is proportional to the abundance of the tRNA species that recognize it (Figure 27, see more details in Subsection 9.5.2).

Denoting the probability that the $i-th$ site is occupied at time t by $p_i(t)$, it follows that the rate of ribosome flow into/out of the system is given by: $\lambda[1-p_1(t)]$ and $\lambda_n p_n(t)$ respectively. The rate of ribosome 'flow' from site i to site $i+1$ is given by: $\lambda_i p_i(t)[1-p_{i+1}(t)]$ (see Subsection 9.5.2). As we discuss in details (see Subsection 9.5.2 and Figure 27D), the RFM and the ASEP, give similar predictions, yet the RFM runs markedly faster.

In this chapter we focus on the steady state solution of the equations presented in Figure 27C and specifically in the rate of protein production at steady state. Steady state is a widely used assumption in cases like these (see, for example, [90, 117, 118]) and is hence a good starting point for a large scale study as the one conducted here. In addition, a pioneering analysis that took into account mRNA degradation and was not based on the steady state assumption, was unable to improve the predictive power of the model with respect to existing data (Subsection 9.5.5). We note however, that this line of investigation is far from being exhausted and that it should be revisited once degradation rates of mRNA molecules and proteins become available (this data is currently lacking for the vast majority of genomes and heterologous genes).

We denote the steady state site occupation probabilities by $\{\pi_1, ..., \pi_n\}$ and the steady state ribosome flow through the system by R. The latter denotes the number of ribosomes passing through a given site per unit time and we note that this rate is nothing but the steady state rate of protein production.

9.2 Basic Properties of the Ribosome Flow Model

One advantage of the RFM is its amenability to both analytical and numerical analysis. In particular one can study ribosome density profiles and protein production rates from the equilibrium dynamics of the translation process. In Subsection 9.5.2 we describes how to solve the model analytically under steady state conditions; in this section we discuss some of the basic properties of the solution.

9.2.1 The behavior of the model under very low and very high initiation rates

A central debate in the field is about the rate limiting stage of gene translation: i.e. is it the initiation stage or the elongation stage (see, for example, [89]). Analysis of our model demonstrates that, in principle, both cases are possible.

As can be seen in Figure 27A, at very low initiation rates, $\lambda \ll min\{\lambda_1, ..., \lambda_n\}$, the initiation rate, λ, is the rate limiting step of the translation process (i.e., it is the bottleneck and the translation rate is determined by it). Thus, the translation rate is approximately given by λ. On the other hand, at high initiation

rates, $\lambda \gg min\{\lambda_1, ..., \lambda_n\}$, the rate limiting step is the elongation ("the flow from codon to codon"); in this case, the rate of protein translation converges to a constant that is determined by the set of elongation rates $\{\lambda_i\}$ (Figure 27A; also see Subsection 9.5.2).

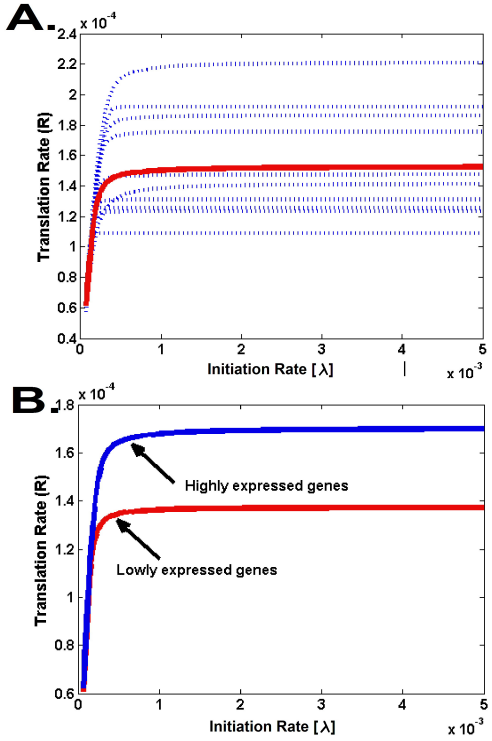

Figure 28: The effect of the initiation rate on the translation rate and elongation rate capacity. A. The figure depicts ten typical profiles of translation rate vs. the initiation rate (blue) in S. cerevisiae genes; the mean genomic profile is shown in red. As can be seen, for very small values all genes have similar translation rates (mainly determined by λ and not by the codon-bias), whereas for larger λ translation rates differ among genes and asymptotically converge to the elongation rate capacity. B. The predicted translation rate for highly (top 25%, Blue line) and lowly (lowest 25%, Red line) expressed genes.

9.2.2 The elongation rate capacity of a coding sequence

One important feature that was discovered by implementing our model is the fact that each gene has a different translation elongation capacity. This capacity is the maximal translation rate of the gene, achievable for infinitely large λ. In effect, one needs not go to "infinitely large" values of λ since the limiting capacity is already achieved for finite and biologically feasible values. As can be seen in

Figure 27A (for large λ), the capacity is a finite number that depends on the mRNA sequence; in addition, for each gene there is a possibly different λ_c, such that for every initiation rate λ above λ_c, the elongation capacity is roughly equal to the maximal elongation capacity. As expected, Figure 27B shows that the elongation rate capacity of highly expressed genes is higher than the capacity of lowly expressed genes (S. cerevisiae; also see Section 9.5).

9.3 Predicting translation rates, protein abundance and ribosome densities of endogenous genes

9.3.1 Translation rates and protein abundance

The model was first evaluated by an analysis of three organisms for which large scale Protein Abundance (PA) measurements are available: E. coli, S. pombe and S. cerevisiae (see Section 9.5). It is important to note that direct measurements of translation rates are not available. However, as explained in Subsection 9.5.7, the protein abundance of a gene is expected to increase monotonically with its translation rate. Thus, a good predictor of translation rates is expected to have a high Spearman correlation with the corresponding protein abundance. Indeed, throughout the paper we mainly report correlation of RFM translations rates with protein abundance (see Section 9.5). We compare the predictions of the RFM to the predictions of other commonly used predictors.

In each case, genes were divided into groups/bins (of equal size) according to their expression levels and the number of protein abundance measurements (a larger number of measurements, e.g. the data of S. cerevisiae, enables more bins); in each group the correlation between the predictions of the model and the actual protein abundance level was computed. The predictions of the RFM are compared with those of the tAI, which is the current state of the art, codon bias based, PA predictor [90, 103, 108, 109, 111, 112, 122]. The RFM and tAI share resemblance in the sense that they are both based on codon adaptation to the tRNA pool. However, in contrast to the RFM, the tAI is not sensitive to the order of codons or to the effect caused by ribosome jamming. The tAI is also a central component in other PA predictors that incorporate additional genomic features such as mRNA levels and evolutionary rates [108]. Thus, whenever the predictions of the RFM are better than those of the tAI, it can beneficially replace the latter as a component within a more sophisticated predictor.

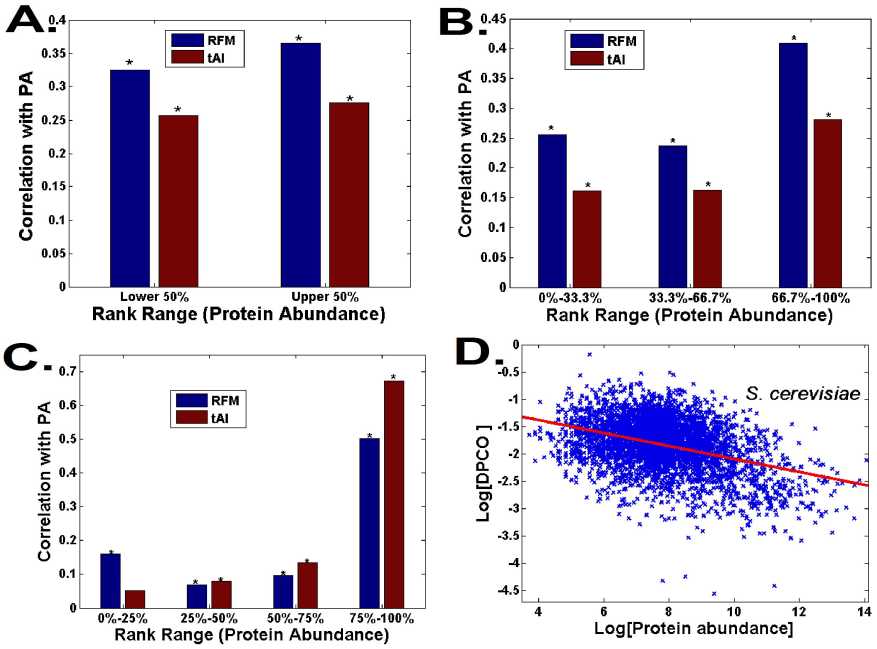

Figure 29: Prediction of protein abundance of endogenous genes by the tAI [109] and by the ribosome flow model (RFM). We compare the RFM to the tAI (insensitive to codon order), the RFM also outperformed other predictors, such as the Bottleneck and the Mean Speed (see definitions in Section 9.5; see Figure 33 in Subsection 9.5). The predictions were obtained for groups of genes with different levels of protein abundance in different organisms; in each organism all bins are of equal size; organisms with a larger number of measurements enable more bins. A. Predicting protein abundance of E. coli endogenous genes. B. Predicting protein abundance for S. pombe endogenous genes C. Predicting protein abundance for S. cerevisiae endogenous genes [99]. D. Sensitivity to codon order vs. protein abundance in S. cerevisiae.

As can be seen in Figure 29, in the vast majority of organisms and across expression levels, the RFM outperforms the tAI (and other predictors that are based on codon bias). Specifically, in E. coli the global correlation between PA and the predictions of the RFM is $R = 0.54$ ($p < 10^{-16}$) vs. $R = 0.43$ ($p < 10^{-16}$) for the tAI (408 genes with PA data). In addition, when subdividing into expression levels, correlations are consistently higher in all subgroups (Figure 29). In S. pombe results were similar: the correlation with PA was higher for the RFM, $R = 0.63$ ($p < 10^{-16}$) vs. $R = 0.56$ ($p < 10^{-16}$) for the tAI (1465 genes with PA data). In addition, correlations are higher in most of the expression level subgroups (Figure 29).

In the case of S. cerevisiae the tAI performs better than the RFM only for the most highly expressed genes. Nevertheless, it is the RFM that yields significant correlation with protein abundance in most of the other ranges (see Figure 29C). This may be due to the tendency of highly expressed genes in S. cerevisiae to be more robust to permutations of the codons' order (see discussion in the next subsection) and due to the fact that the tAI was specifically tailored and optimized for S. cerevisiae [109].

Finally, the RFM is seen to outperform the tAI also when mRNA levels are controlled for and when the product of the predicted translation rate with the mRNA level of the transcript is used as the PA predictor; see Subsection 9.5.19.

9.3.2 The effect of codon order on translation rates

All common measures of translation rate/translation efficiency/codon bias (see, for example, [109, 110]) predict that PA increases with the relative incidence of 'fast' codons along the transcript. Recently, it has been suggested that codon order (in addition to content) may regulate gene translation via the effect of ribosome jamming [90, 105, 106]. For example, slower codons at the end of the mRNA, may render the transcript prone to more 'traffic jams' and thus decrease the translation rate. Previous studies have attempted to estimate the effect of codon bias in the case were synonymous codons are randomly permuted and the final protein product does not change [89, 104]. Nevertheless, common measures of translation rate are not sensitive to codon order and so a direct estimation regarding the effect of the latter on the translation rate is still lacking.

In this section, we aim at isolating the effect of codon order on the translation rate. In other words we would like to answer the following question: is there a difference between the translation rates of two mRNA transcripts that are characterized by identical codon content but different codon order. To this end, we applied our model to random permutations of native mRNA transcripts. This was done for each gene separately, in order to compute the standard deviation in the predicted PA for the set of randomly permuted transcripts. Results are given in percentages (i.e. normalized by the original PA; see the exact details in Subsection 9.5.13). We named this measure DPCO (dependence of protein abundance on codon order). We emphasize again that DPCO analysis cannot be performed using common measures of translation rate/translation efficiency since these are only sensitive to the codon content which was left unchanged by the permutation process.

A DPCO index of 20%, for example, means that we can quite easily get a 20% change in the gene's PA just by changing the order of its codons, and probably get a 40% change in PA by optimizing the latter with respect to codon order. Codon permutations may change the resultant protein; nevertheless, the DPCO gives a large scale estimation of the distinct effect of codon order on protein production rates and protein abundance.

Analysis of several organisms revealed that the DPCO of endogenous genes is surprisingly high. The mean DPCO is 16.35% in E. coli (stdev is 8.43%: in 10% of the genes the DPCO is more than 28%); the mean DPCO is 13.7% in S.

pombe (stdev of is 4.6%: in 10% of the genes the DPCO is more than 19.25%); the mean DPCO is 17.7% in S. cerevisiae (stdev 7.92 %: in 10% of the genes the DPCO is more than 27.46%). These results highlight the importance of incorporating codon order into models of translation rates as they support the hypothesis that one can profoundly affect the translation rate just by reordering the codons in the transcript.

In the previous section we found that the tAI performs well mainly for highly expressed genes; it is possible that this result is partially related to the fact that translation efficiency is less affected by codon order in these genes. We found a significant negative correlation (S. cerevisiae: $r = -0.31$, $p < 10^{-16}$; E. coli: $r = -0.22$, $p = 9.4 \cdot 10^{-6}$) between DPCO and protein abundance of genes (Figure 29D), demonstrating that in these organisms protein abundance of highly expressed genes (whose expression was predicted relatively well by the tAI) is less dependent on codon order than it is in lowly expressed genes. Thus, the results reported in this section support the usage of models such as the RFM for predicting the translation rate of endogenous genes that are lowly expressed (see also Subsection 9.5.20).

It is important to note that the predictions reported in this section should be confronted with experimental measurement when these become available. However, in light of the fact that controlled design of 'wet experiments', that would allow the validation of the predictions presented above, is far from being trivial (e.g. changing the order of codons may influence other features of the coding sequence), the estimations reported here are particularly interesting.

9.3.3 Coarse graining and genomic ribosomal density profiles

Figure 30A depicts the correlation between translation rate predictions of our model and protein abundance in S. cerevisiae for different values of the coarse graining parameter C (C in Figure 27). Interestingly, the optimal correlation is obtained for sites of size 25-35 codons (and is supported by jackknifing test; see Subsection 9.5.17). This value is similar to length scales associated with the ribosome such as its footprint on the mRNA sequence [90, 97, 114, 115, 116] (between 11 and 18 codons), the number of amino acids associated with the exit channel of the ribosome and its length [123, 124, 125, 126] (between 30 and 71 codons), and the length of the 'ramp' at the beginning of genes corresponding to the optimization of ribosome allocation [90, 125] (around 50 codons); similar results were obtained for other organisms as well (Figures 34-35 in Subsection 9.5.24). This result provides further support for the validity of our model. Specifically, this result is consistent with the assumption that site size in our model should be of the same order of magnitude as the ribosome size since physically this is the relevant length scale in the system.

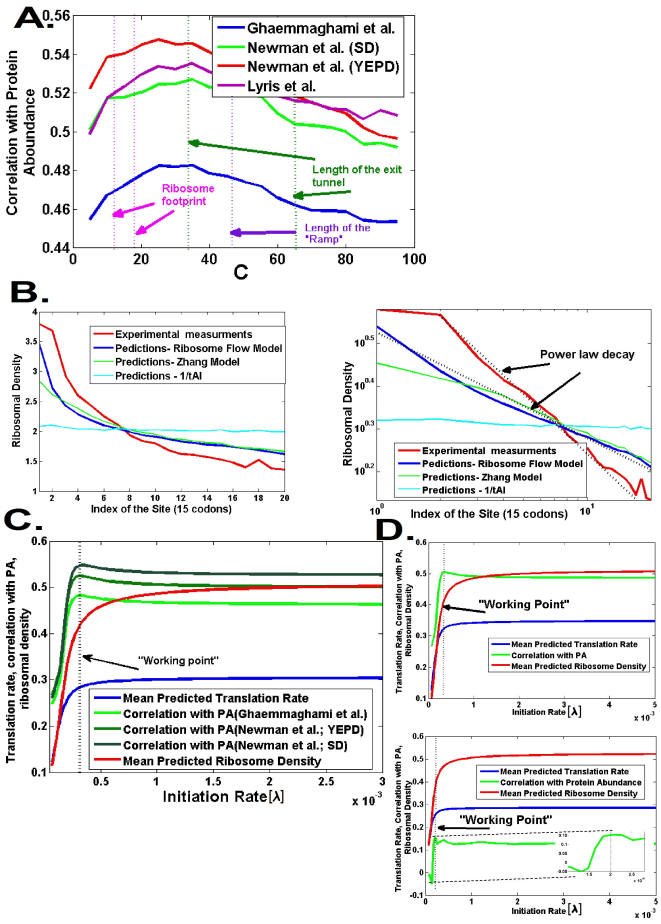

Figure 30: Relations between various quantities predicted by the RFM and biological measurements. A. Correlation between protein abundance [98, 99, 147] and the translation rate for various values of the coarse graining parameter (C in Figure 2); the best results are observed for values which are similar to various geometrical properties of the ribosome (the dashed lines in the figure). B. Right: The RFM predicts the genomic ribosomal density profile [14] better than the tAI or the model of Zahng et al. [114]; all were normalized to have the same mean. Left: the 5' region of the genomic ribosomal density profile and the predicted genomic profile of the RFM appear linear on a log-log scale. We used a site size of 15 codons (similar to the size of the ribosome) and a (initiation rate) value that was indepedently found to optimize the correlation with protein abundance. C. The relation between λ (associated with the number of available ribosomes in the cell), genomic mean of the translation rate, and the genomic mean of the ribosomal density. D. Initiation rate λ, translation rate, and ribosomal density for highly expressed genes (up) and lowly expressed genes (down).

In the next step, we studied how well the RFM predicts the shape of the genomic profiles of ribosome density. To this end, predictions of our model and other models were compared to a genomic ribosomal density profile that was generated based on, single nucleotide resolution, large scale measurements of ribosomal density (see Section 9.5, [97]; Figure 30B).

Strikingly, as depicted in Figure 30B, although all models predict that there is a decrease in ribosome density from the 5' end to the 3' end of the mRNA transcript, the gap between the real profile of ribosomal density and the profile predicted by the RFM is significantly smaller than the one obtained by Zhang's model [114] (0.26 vs. 0.3; Wilcoxon test p-value < 0.0001) or from the graph corresponding to per-codon mean genomic 1/tAI [90] (0.26 vs. 0.54; Wilcoxon p-value < 0.0001). Specifically, it seems that both the genomic ribosome density profile and the RFM predictions are characterized by a non-exponential decay from the 5' end of the coding sequence to the 3' end of the coding sequence and are seen linear on a log-log graph (Figure 30B; see also Subsection 9.5.21). In contrast, the tAI predicts a much slower mean genomic decrease rate (Figure 30B). This result further supports the RFM as a model that describes the physics of gene translation better than previously suggested models (similar results were obtained for ribosome density profiles obtained under starvation conditions; see Figure 36 in Subsection 9.5.24).

9.3.4 Optimality of the translation machinery

One basic translation-related feature of a gene is the mean, steady state, ribosome density on the transcript. This value can be predicted by $\bar{\pi} = \frac{1}{N}\sum_{i=1}^{N}\pi_i$ (the mean probability that a site will be occupied by a ribosome). In the RFM, λ models the effect of the number of free ribosomes on the initiation rate. Given that there are more ribosomes, the initiation rate would increase since the rate in which ribosomes arrive at the 5' end of the mRNA is proportional to the number of free ribosomes. What are the relations between $\bar{\pi}$, λ, and the translation rate in general? And in particular, what is the actual 'working point' (in the λ, $\bar{\pi}$, R parameter space) of the translational machinery?

Figure 30C depicts the translation efficiency at different values of λ. At low λ levels the translation rate and ribosome occupancy increase monotonically with λ. However, as was demonstrated before [127], after a certain point the system reaches saturation — increasing λ does not result in a further increase of the translation rate or the mean genomic ribosomal density.

Interestingly, the correlation between the predicted translation rate and the measured protein abundance of yeast is maximal exactly before the onset of saturation (Figure 30C). This fact may suggest that the translation machinery is tuned to work in the vicinity of this point. Thus, this may indicate that there is global optimality of the initiation rate in S. cerevisiae (similar results were obtained for other organism: S. pombe, E. coli, Human liver; see Figures 37-39 in Subsection 9.5.24).

We note that the pre-saturation point is optimal from an engineer's point

of view. The basic reasoning for this follows from the fact that going below the pre-saturation dramatically decreases the rate of protein production. On the other hand, going above and beyond the pre-saturation point, would require additional resources from the cell. This investment however, will have no effect on the mean protein production capacity and will therefore be in vein.

For a given initiation rate, λ, faster codons (i.e., higher λ_i or higher tAI) should decrease the ribosomal density due to the reciprocal relation between translation rate and ribosomal density [90, 103]. Thus, under the assumption of a global initiation rate, and since highly expressed genes have more efficient codons, we expect a negative correlation between expression levels of genes and their ribosomal density. However, in practice this is not the case — the correlation between translation efficiency (tAI) and ribosomal density is positive and significant (for example, $r = 0.46$; $p \leq 10^{-16}$ for the ribosomal density measurements of [93] and the mRNA measurements of [128]). This result suggests that the initiation rate of highly expressed genes is higher than that of lowly expressed genes. Refining our analysis, we will now revisit, and relax, the simplifying global initiation rate assumption we have made so far.

Given a set of genes (e.g. highly expressed genes) the estimated initiation rate λ of this group is the one that gives the best correlation between the predicted translation rates and protein abundance. We estimated the initiation rate in highly expressed genes (top 20%) and in lowly expressed genes (lowest 20%; Figure 30D). Indeed the predicted initiation rate of the highly expressed genes is higher than that of the lowly expressed genes (0.00035 vs. 0.0002) while the resulting predicted ribosome density is also higher for the highly expressed genes (0.42 vs. 0.36). Thus, in practice (at the 'working point'), our model predicts that highly expressed genes, that are equipped by faster codons and thus characterized by higher translation rates, are also characterized by higher ribosomal densities as their initiation rate is higher. The fact that in highly expressed genes ribosomal densities are higher, suggests that in these genes, elongation rate is more rate-limiting (relatively to lowly expressed genes). This result explains why in highly expressed genes codon bias should be a better predictor of translation rate (as was shown in Figure 29).

Different mRNA transcripts are characterized by different translation elongation capacities. Here, based on the correlation between translation rates and protein abundance, we have just shown that, on average, the predicted λ is the one for which this capacity is almost fully achieved (i.e. 93% of the capacity is attained in S. cerevisiae). This rule enables inference of the initiation rates of individual genes: e.g. in S. cerevisiae, the predicted initiation rate of a gene is the one for which 93% of its elongation capacity is attained (in other organisms the rule is similar; see Section 9.5).

Strikingly, the predicted initiation rate of genes significantly correlates with their protein abundance (S. cerevisiae $r = 0.29$, $p = 10^{-16}$; S. pombe $r = 0.41$, $p = 10^{-16}$; E. coli, $r = 0.34$, $p = 8 \cdot 10^{-13}$, Figures 40-42 in Subsection 9.5.24); i.e. highly expressed genes have higher initiation rates. In addition, the predicted initiation rate correlates with the predicted ribosomal density (S. cerevisiae $r = 0.72$, $p < 10^{-16}$; S. pombe $r = 0.6531$, $p < 10^{-16}$; E. coli,

$r = 0.3379$, $p < 10^{-16}$, Figures 43-45 in Subsection 9.5.24, Section 9.5) — i.e. highly expressed genes are characterized by higher ribosomal density (the correlation between predicted ribosome density and protein abundance of genes: S. cerevisiae $r = 0.19$, $p < 10^{-16}$; S. pombe $r = 0.104$, $p = 2.44 \cdot 10^{-4}$; E. coli, $r = 0.32$, $p = 2.1 \cdot 10^{-11}$; Figures 46-48 in Subsection 9.5.24) . These results demonstrate again that the predictions of our model are in accord with the experimental observation that highly expressed genes have higher initiation rate and higher ribosomal density (mentioned above) [93].

9.3.5 Analysis of heterologous gene expression

As was demonstrated above, the RFM is considerably better (than current state of the art predictors) at predicting the PA of lowly expressed genes with coding sequences that differ from the optimal design. This is usually the case when a gene from one organism (e.g. Human) is expressed in a different organism (e.g. E. coli; see for example, [88, 89, 104, 129]), a procedure known as heterologous gene expression. Heterologous gene expression allows the use of mRNA 'libraries' that are composed of different variants of the same heterologous gene. In this method of expression, control for various properties is already 'built in'. In particular, the amino acids composition of the translated protein remains unchanged.

In this subsection, we use our model to analyzing two cases of heterologous gene expression, demonstrating that the RFM markedly outperforms the tAI (and other alternative predictors). In what follows, we emphasize the differences between endogenous and heterologous genes. As we demonstrate, the gap between the predictions of our model and those of the tAI is higher for heterologous genes. This property of the RFM, demonstrates the potential biotechnological applications of our approach — predicting the protein abundance of heterologous gene expression.

We analyzed the data of Welch et al. [104], a large library of genes encoding DNA polymerase of Bacillus phage pi29 proteins, results are shown in Figure 31. All the genes encode the same amino acid sequence but each of them has a different codon composition. Although it was reported that there is no correlation between codon-bias or folding energy and protein abundance in this dataset [103, 104], we found a significant correlation between the predictions of the RFM and protein abundance ($r = 0.5$, $p = 0.004$). Correlation is significant only for very low initiation rates, suggesting that initiation (or other variable, as was suggested in [104]) is rate limiting in the translation of these genes. In contrast to what was observed for endogenous genes (Figure 30), the point with maximal correlation between the prediction of the model and PA is not the pre-saturation point. This result demonstrates that the coupling between translation rate and initiation rate is an evolutionarily selected trait, and is hence not observed in heterologous coding sequences.

We continued with an analysis of the data by Burgess-Brown et al. who optimized the codons of 31 human genes in order to express them in E. coli [129]. In this study, the protein abundance of 18 genes improved, that of one

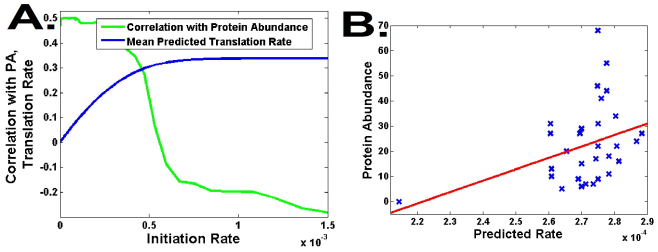

Figure 31: Analysis of the data of Welch et al. [104] by the RFM model. A. The translation rate and the correlation with protein abundance as a function of the initiation rate. B. Predictions of the RFM vs. protein abundance — a dot plot.

gene decreased, and the other 12 did not change in a detectable way. The Spearman correlation between the direction of the change in PA and the predicted fold change (i.e. the ratio between the translation rate before and after the optimization) of the RFM was 0.45 (empirical p-value = 0.019) while the correlation with the fold change according to the tAI was only 0.34 (empirical p-value = 0.077; see Subsection 9.5.15). This result demonstrates once more that the RFM is a particularly useful tool for the analysis of heterologous gene expression (see also Subsection 9.5.22).

9.3.6 Condition-specific translation rates in S. cerevisiae

When the yeast S. cerevisiae is grown on glucose-based media, it first utilizes the available glucose, growing by fermentation. When most of the glucose has been consumed it undergoes a metabolic change, called diauxic shift, in which its metabolism shifts to respiration. This is accompanied by wide changes in gene expression and tRNA abundance [90, 130]. In [90] we focused on the similarities between the tRNA pools in different stages of the diauxic shift (for example, the Spearman correlation between the tRNA abundance at time 0 and the tRNA abundance after 9 hours is 0.9, p-value $6 * 10^{-15}$; i.e. 0.81 of the variance in the tRNA pool at time 9 hours can be explained by the tRNA pool at time 0 hours). In the current study we analyze the dissimilarities between the tRNA pools during different stages of the diauxic shift. Changes in the tRNA pool due to the diauxic shift lead to changes in the translation rate of different codons. The total effect of these changes is related, among other factors, to the order of codons along the mRNA transcript and therefore cannot be inferred completely by the tAI.

Here, we use our model to analyze the dynamics of genomic translation rates during the diauxic shift in S. cerevisiae (using data from [90]). In each stage of the diauxic shift, we computed the expected translation time (t_i) of each codon based on the available tRNA pool at that stage [90]. These times where

then used in conjugation with the RFM in order to compute the mean genomic translation rate and ribosomal densities for different values of the initiation rate λ.

As the new growth conditions are less optimal for the yeast we expect a global reduction in the rate of translation. The mean genomic profile of the translation rate and ribosomal density of all S. cerevisiae genes at five time points (0, 4.5, 6, 7.5 and 9 hours after the beginning of the experiment) during the diauxic shift, is presented in Figure 32A-B. As can be seen, all these profiles are similar to the ones reported earlier — displaying saturation of the translation rate and the ribosomal density for large λ.

As expected, both the predicted translation rate and the predicted number of available free ribosomes (or equivalently the initiation rate) decrease during this process (Figure 32A). Interestingly, although the mean codon efficiency remains essentially unchanged during the process (a minor decreased of 0.16% in the mean genomic expected time for translating a codon), the mean production rate does decreases due to changes in the initiation rate (number of free ribosomes; see details in Figure 32A) and effects related to the flow of ribosomes and the order of codons. In contrast, the mean predicted ribosomal density does not decrease as λ decreases (see details in Figure 32B). Thus, while the total effect under these conditions is also related to changes in mRNA levels, initiation/elongation factors and more (see [130]), our model predicts that part of the global response can be attributed to changes in the composition of the tRNA pool. Such an analysis cannot be performed by simple measures such as tAI.

In the next step, we checked how well the predicted change in translation rate of genes during the Diauxic shift correlates with the change in their mRNA levels. We compared the change in the predicted translation rate of genes whose mRNA levels exhibited extreme fold change (fold changes >1.8 and $<1/1.8$) and found that the ranked fold changes of the translation rate of the genes in these groups was also significantly different (mean fold change 1.035 vs. mean fold change 0.9991; $p = 2.47 \cdot 10^{-5}$). Ranking the changes in the tAI led to an opposite result — a decrease in the translation rate of genes whose mRNA level increased and vice versa (mean fold change 0.9923 vs. mean fold change 1.0103), demonstrating again the superiority of our model. This result demonstrates that (i) in S. cerevisiae, condition-specific changes in the translation rate of genes are in accordance with the changes in their transcription levels; and (ii) the RFM, by considering refined features such as the order of codons and initiation rates is specifically sensitive to the adaptation of an organism to a dynamically changing environment.

9.3.7 Translation Efficiency in Human

Finally, comparison of the predictions of the RFM to tissue specific mRNA levels (that are known to correlate with protein abundance and ribosomal densities [93, 97, 108]) in human demonstrated that it outperforms the tAI in this organism as well (Figure 32C, Subsection 9.5.23). Specifically, the gap between the RFM

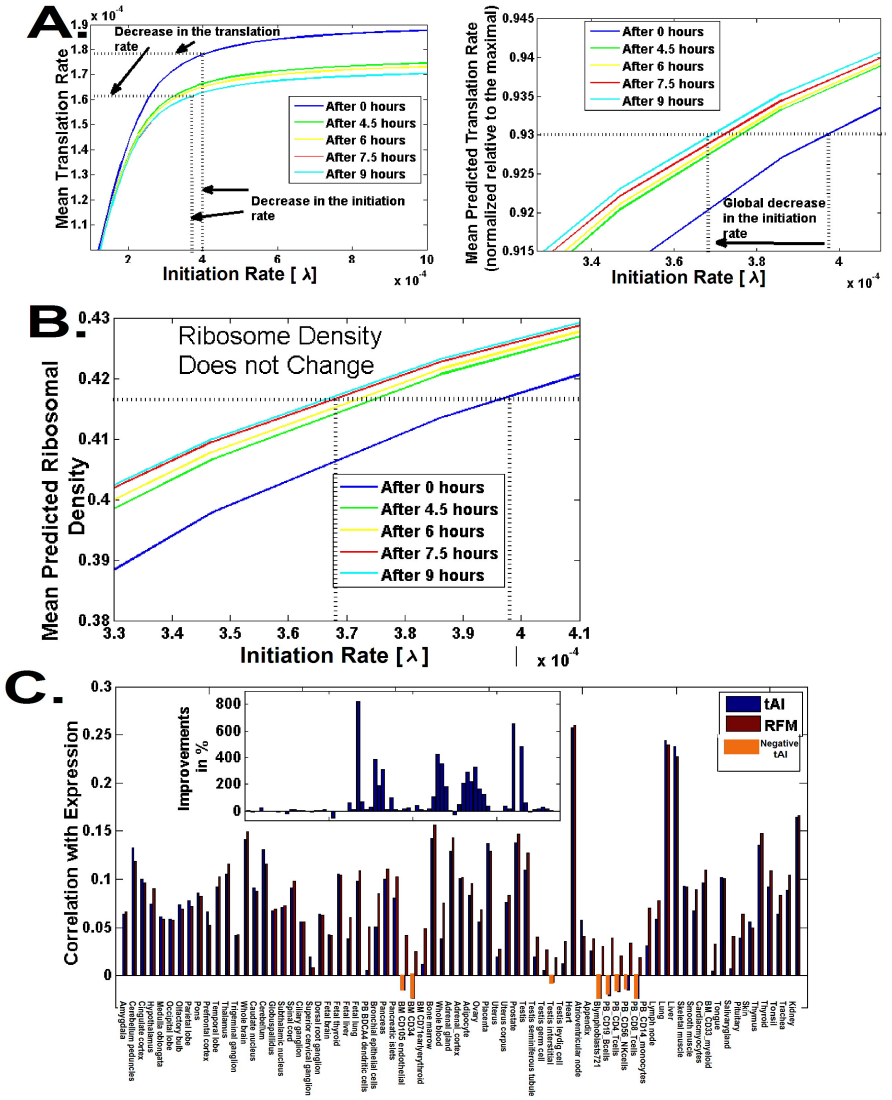

Figure 32: Translation rate and ribosome density during the diauxic shift in S. cerevisiae. A. The mean genomic translation rate as a function of the initiation rate (λ) for five time points; the dotted lines correspond to the working point just before saturation (93% of the maximal production rate). B. The mean genomic ribosomal density as a function of the initiation rate (λ) for five time points. The dotted lines correspond to the initiation rates at the working points. C. The correlation between the mRNA levels of genes in different human tissues vs. (a) the RFM predictions and (b) the tAI predictions. Inset: the improvement in correlation in % when using the RFM instead of the tAI.

and the tAI is particularly large in germ line and immune cell types. Thus, specifically in these tissues, the RFM should be helpful in analyzing mutations (see, for example [122]) or SNPs (see, for example, [85, 131, 132]) that cause diseases due to problems in gene translation.

In addition, we computed the correlation between the prediction of the RFM and protein abundance in Human cell lines for which PA data exists [133]. The correlation between the predictions of the RFM and protein abundance was 0.47 (p-value $< 10^{-16}$) vs. a correlation of only 0.28 (p-value $< 10^{-16}$) between the tAI and protein abundance.

9.4 Conclusion and Discussion

We described a novel analysis of large scale genomic data by a predictor/model that is based on the physical and dynamical nature of gene translation. Given the copy numbers of the tRNA genes in the host genome, our model, the RFM, is based only on codon-bias; It can hence be applied when only the coding sequence of a gene is available and without additional data or information. Despite its relative simplicity, we show that our model predicts features such as protein abundance in endogenous and heterologous genes better than alternative ('non-physical') approaches. We demonstrate that the gap between the performance of the RFM and alternative predictors is especially large in the case of heterologous genes; thus, it should be very helpful in the common challenge of predicting the protein abundance of potential heterologous proteins before expressing them in the desired host (see, for example, [88, 89, 90, 104, 134, 135, 136, 137]). In addition, we have demonstrated that our approach can be used for accurately inferring various variables that can not be inferred by the common predictors used nowadays.

From a Systems Biology point of view, by using our model we were able to demonstrate the global optimality of the process of gene translation [89, 90, 103]. We discovered that increasing the number of available ribosomes (or the initiation rate) increases the genomic translation rate and the mean ribosomal density only up to a certain point. After this point, the system is 'saturated': adding more ribosomes/increasing the initiation rate does not result in an increase of these two variables. Quite strikingly, in all the organisms we have analyzed, the global initiation rate is optimized to the pre-saturation point. The fact that similar results were not observed in artificial genes supports the conclusion that this feature is under selection.

Optimality of the translation machinery is perhaps not so surprising. Protein production is a central and complex process in the cell. For example, at any given time point there are around 60,000 mRNA molecules in S. cerevisiae [117] that are translated by 187,000 (+56,000) ribosomes [118]. The process of gene translation consumes a very large amount of energy and thus the problem of fine tuning the number of ribosomes and the translation rate should have a significant influence on the fitness of the organisms [89, 90, 103]. Specifically, increasing the translation rate of highly expressed genes (the 'supply') while decreasing the number of working ribosomes/ribosomal density (the 'cost') should

improve the fitness of an organism. It was already suggested that there is selection for improving translation efficiency of highly expressed genes relatively to lowly expressed genes (see, for example, [89, 103]). By using our model, we can actually estimate the translation cost of highly and lowly expressed genes as the ratio between the translation rate and the average number of ribosomes working on the transcript. The number of proteins produced per unit time, per ribosome, for highly expressed genes (top 20%) is $0.000162/0.42 = 0.000386$ (in arbitrary units). This number is 10% higher than that of the lowly expressed genes (lower 20%; $0.000125/0.36 = 0.000347$). Again, this result demonstrates 'optimality': as highly expressed genes produce more mRNA molecules, decreasing the cost of translation should result in a much larger effect on the fitness of the organism.

Finally, the goal of this study was to model the process of translation elongation, emphasizing the effect of codon order. In the future, in order to decrease the gap between the predictions of our models and measurements of protein abundance, we intend to develop a more comprehensive model of this process. While promising strides in this direction were already made [138, 139], many features of the translation process are yet to be accounted for. Unfortunately, large-scale biological measurements of translation rates, initiation rates, tRNA levels, mRNA/protein degradation rates and many other quantities that are related to the process of gene translation are currently unavailable. Large scale measurements that are available (e.g. protein abundance) are related to the modeled process (see Section 9.5), but are indirect. This fact hinders the implementation and validation (as opposed to formulation) of more sophisticated models. In addition, it is important to note that the ability to predict measurements of protein abundance may also be hindered due to bias and noise in the current pool of existing data (see, for example, [100, 140]). As new data accumulates, the implementation of more comprehensive models will become possible and our understanding of the translation process will deepen further.

9.5 Appendix

9.5.1 The ASEP model for translation elongation

In the ASEP an mRNA transcript with N codons is modeled as a chain of sites, each of which is labeled by the index i, where $i = 1, ..., N$. The first and last codons are associated with the start and stop codons, respectively. At any time, t, attached to the mRNA are $M(t)$ ribosomes. Being a large complex of molecules, each ribosome will cover l codons. A codon may be covered by no more than a single ribosome. To locate a ribosome, we arbitrarily assume that the codon being translated is the one in the 'middle' of the ribosome. For example, if the first, $(l+1)/2$ codons are not covered, a ribosome can bind to the first codon on the mRNA strand, and then it is said to be "on codon $i = 1$". A complete specification of the configuration of the mRNA strand is given by the codon occupation numbers: $n_i = 1$ if codon i is being translated and $n_i = 0$ otherwise. Note that when $n_i = 1$ the $(l-1)/2$ codons before and after codon $i = 1$ are covered by the ribosome that is on site i. Since these codons are not the ones being translated, the codon occupations numbers for them are equal to zero.

We will now specify the dynamics of the ASEP model. A free ribosome will attach to codon $i = 1$ with rate λ, provided that the first $(l+1)/2$ codons on the mRNA are empty. An attached ribosome located at codon i will move to the next codon with rate λ_i, provided codon $i + (l+1)/2$ is not covered by another ribosome. In case $i + (l+1)/2 > N$ (ribosome is bulging out of the mRNA strand) an attached ribosome will move to the next codon with rate λ_i.

In order to simulate this dynamics, we assume that the time between initiation attempts is distributed exponentially with rate λ. Similarly the time between jump attempts from site to site is assumed to be exponentially distributed with rate λ_i (the exponential distribution is of course, an approximation as the process of translating a single codon involves more than one step [84]). Note that in the case of $i = N$ the jump attempt is in fact a termination step. We define an "event" as an initiation, jump attempt, or termination step. From our definition it follows that the time between events is exponentially distributed (minimum of exponentially distributed random variables) with rate $\mu(\{n_i\}) = \lambda + \sum_{i=1}^{N} n_i \lambda_i$. Note that a jump attempt from codon i can only be made if there is a ribosome translating this codon and hence the rate depends on the set of site occupation numbers. The probability that a specific event was an initiation attempt is given by: $\lambda/\mu(\{n_i\})$. Similarly, the probability that a specific event was a jump attempt (or termination event) from site to site is given by $n_i\lambda_i/\mu(\{n_i\})$.

At each step of the simulation, we determine the nature of the event and the time passed till its occurrence by these rules. The set of site occupation numbers are then updated accordingly and the simulation proceeds to the next event. For example if an initiation attempt was made, we check if the first $(l+1)/2$ codons on the mRNA are not covered. If so, we set $n_1 = 1$, otherwise

the attempt fails and n_1 remains as is. If a jump attempt from codon i to codon $i+1$ was made, we check if site $i+(l+1)/2$ is not covered. If so, we set $n_i = 0$ and $n_{i+1} = 1$, otherwise the attempt fails and n_i,n_{i+1} remain as is. Starting with an empty mRNA strand we simulate the system for 250,000 steps (events). The system is then simulated for an additional 1,000,000 steps where we keep track of the total number of terminations and the total time that have passed from the point this phase have started. The steady state rate of protein production was determined by dividing the number of termination events by the total time that has passed. The number of steps in the first and second stages was determined after observing that increasing the number of steps fourfold had a negligible effect on the predicted protein production rate.

9.5.2 The Ribosome Flow Model

Physical interpretation of the ribosome flow model. Assume that a ribosome is C condos long and that the mRNA strand is positioned such that translation takes place from left to right. The ribosome flow model assumes that a ribosome lands on the mRNA strand such that the first codon is located at the middle of the ribosome. The ribosome now needs to translate C codons in order to have its middle point reach codon C+1. This way the right edge of a newly arriving ribosome can be positioned next to the left edge of the ribosome who has just translated the first C codons. We now coarse grain the mRNA strand into two groups of sites ('chucks'):

- A. $1 \ldots (C+1)/2, 1+(C+1)/2 \ldots C+(C+1)/2, 1+C+(C+1)/2 \ldots 2C+(C+1)/2, \ldots$
- B. $1 \ldots C, C+1 \ldots 2C, 2C+1 \ldots 3C, \ldots$

The flow of ribosomes from site i to site $i+1$ in the group A is determined by:

1. The occupation probabilities of these sites. The higher the occupation probability of site i (more attempts per unit time to flow from site i to site $i+1$) the higher the flow to site $i+1$. The higher the occupation probability of site $i+1$ (more chances that a ribosome will be blocked by another ribosome residing in site $i+1$ when attempting to flow from site i to site $i+1$) the lower the flow emanating from site i.

2. The translation time of the C codons that belong to the $i-th$ site in group B. The lower the time the higher the flow.

Denoting by $p_i(t)$ the probability that the $i-th$ site (in group A) is occupied (by a ribosome) at time t, the time evolution of the set of probabilities $\{p_1(t), ..., p_n(t)\}$ is governed by the following set of differential equations:

$$\begin{cases} \frac{dp_1(t)}{dt} = \lambda \left[1 - p_1(t)\right] - \lambda_1 p_1(t) \left[1 - p_2(t)\right] \\ \frac{dp_i(t)}{dt} = \lambda_{i-1} p_{i-1}(t) \left[1 - p_i(t)\right] - \lambda_i p_i(t) \left[1 - p_{i+1}(t)\right] \quad 1 < i < n \\ \frac{dp_n(t)}{dt} = \lambda_{n-1} p_{n-1}(t) \left[1 - p_n(t)\right] - \lambda_n p_n(t) \, . \end{cases} \tag{338}$$

Analytic solution of the ribosome flow model. In order to proceed we recall that in steady state the occupation probabilities are constant in time and equal to $\{\pi_1, ..., \pi_n\}$. Denoting the steady state rate of protein production by R it follows that:

$$R = \lambda_n \pi_n \tag{339}$$

This rate is also equal to the steady state rate at which ribosomes leave the mRNA strand (after translating the entire sequence). At steady state the left hand side of Eq. (338) vanishes and we get:

$$\begin{cases} \lambda \left[1 - \pi_1\right] = \lambda_1 \pi_1 \left[1 - \pi_2\right] = R \\ \lambda_{i-1} \pi_{i-1} \left[1 - \pi_i\right] = \lambda_i \pi_i \left[1 - \pi_{i+1}\right] = R \quad 1 < i < n \\ \lambda_{n-1} \pi_{n-1} \left[1 - \pi_n\right] = \lambda_n \pi_n = R \,. \end{cases} \tag{340}$$

where we have also used Eq. (339). An interesting conclusion follows from Eq. (340), since for every site i: $0 \leq \pi_i \leq 1$ (probability is always non-negative and not larger than one) the steady state rate of protein production is limited by the slowest rate in the system:

$$R \leq min \left\{\lambda, \lambda_1, ..., \lambda_n\right\} \tag{341}$$

Solving Eq. (340) for R we obtain:

$$1 - R/\lambda = \cfrac{R/\mu_1}{1 - \cfrac{R/\mu_2}{1 - \cfrac{R/\mu_3}{1 - \cfrac{R/\mu_4}{1 - \cfrac{R/\mu_5}{\ddots}}}}} \tag{342}$$

Equation (342) is the starting point for the analytical analysis of the model as is further described below. Note that in principle Eq. (342) can be solved numerically for R given the set $\{\lambda, \lambda_1, ..., \lambda_n\}$, the unknown steady state occupation probabilities $\{\pi_1, ..., \pi_n\}$ can then be computed via Eq. (340). In practice however, we have numerically solved the original set of differential equations given in Eq. (338).

Solving Eq. (338) numerically

In order to obtain the set of steady state occupation probabilities, $\{\pi_1, ..., \pi_n\}$, and the steady state rate of protein production, R, we solve Eq. (338) numerically using Matlab. Equation (338) is treated as an ordinary differential equation for the vector $\vec{p}(t)$ whose entries are the occupation probabilities: $\{p_1(t), ..., p_n(t)\}$. We start from an mRNA strand which is empty of ribosomes, $\vec{p}(t) = \vec{0}$. The occupation probabilities are then found for a set of later times using Eq. (338) and Matlab's ordinary differential equation solver. The process stops when the vector converges to the vector of steady state occupation probabilities. More accurately, we stop the process for a time t^* for which $\vec{p}(t)$ is constant (up to some prefixed numeric error threshold) for every $t > t^*$. The vector of steady state occupation probabilities and the protein production rate are then taken as: $\vec{\pi} = \vec{p}(t^*)$ and $R = \lambda_n \pi_n$.

Analysis of the limits of low and high initiation rates

An interesting question goes to the behavior of the model in the limits of low/high external ribosome flux. The limit of low ribosome flux is mathematically given by: $\lambda \ll min\{\lambda, \lambda_1, ..., \lambda_n\}$. In this limit the rate of protein production may be approximated by $R \simeq \lambda$ and it is hence insensitive to codon bias. In other words, the genomic rate of translation is equal to the rate of ribosome arrival since the latter is the rate limiting step of the process. In order to derive this result we first note that in this limit $R \leq \lambda \ll min\{\lambda, \lambda_1, ..., \lambda_n\}$ by use of Eq. (341). It follows that $R/\lambda_1 \ll 1$ and we may hence approximate by neglecting the right hand side of Eq. (342). The requested result then follows as is further illustrated in Figure 28A.

The limit of high ribosome flux is mathematically given by: $\lambda \gg max\{\lambda, \lambda_1, ..., \lambda_n\}$. In this limit the rate of protein production converges to a transcript specific constant $R^*(\lambda_1, ..., \lambda_n)$ that does not depend on the ribosome flux λ (Figure 28A). Under these circumstances the rate of protein production is strongly affected by codon composition and codon arrangement along the mRNA molecule. In addition, the independence of R on λ implies that above a certain threshold any attempt to increase R by increasing λ is futile. Since increasing λ comes with the cost of spending valuable resources on maintaining a large ribosome pool cost/benefit considerations will set a clear physiological upper bound on λ (see also Subsection 9.3.4). In order to understand the behavior of the protein production rate in this limit we first note that $\lambda \gg max\{\lambda, \lambda_1, ..., \lambda_n\} \geq min\{\lambda, \lambda_1, ..., \lambda_n\} \geq R$ by use of Eq. (341). It follows that $R/\lambda \ll 1$ and we may hence approximate by neglecting this term in the left hand side of Eq. (342). We now see that R is a solution to an equation that does not contain the ribosome flux λ as was argued above. This result is further illustrated in Figure 28.

9.5.3 The ASEP model vs. the RFM

The generalized ASEP model mentioned above is a generalization (elongated particles and site dependent rates) of a simpler ASEP model (see, for example, [141]). In the case of the ribosome flow model, we make two approximations. The first is coarse graining (dividing into chunks/sites), this approximation is quite common and was applied to various physical and biophysical problems. The second approximation is nothing but the mean field approximation. This means that in order to write the master equation for our model (Figure 27C) we have implicitly neglected the fact that there could be correlations between sites. We hence write approximate equations for the average (over many identical mRNA systems) occupation probabilities. Doing so, we assume that the probability that site i is occupied/empty and that site $i+1$ is occupied/empty is well approximated by the probability that site i is occupied/empty times the probability that site $i+1$ is occupied/empty. Although in general this is not always true, this approximation is also common in the ASEP literature.

9.5.4 RFM with abortions

Within the framework of the RFM, abortions were modeled by adding an abortion probability to the model. The abortion probability determines the percent of ribosome-ribosome collisions that will result in abortion, i.e., in premature detachment of the ribosome from the mRNA strand. Mathematically, abortion adds the following term to the model: $-p_{ab} \cdot p_i(t) \cdot p_{i+1}(t)$ where p_{ab} is the abortion probability. For every $1 \leq i \leq N$ this term is added to the $i-th$ and $(i+1)-th$ rows of equation 338. This modification of the RFM corresponds to mutual abortion, i.e., to a situation where after an abortive collision both ribosomes will stop processing the mRNA transcript. Scanning different values for p_{ab}, we discovered that maximal correlations were obtained in the case of $p_{ab} = 0$, i.e. in the limit were abortions due to ribosome-ribosome collisions are negligible.

9.5.5 mRNA half life – steady state revisited

In order to examine the steady state assumption (within the limitations of existing data), we analyzed the RFM model without it. Analysis was performed on the S. cerevisiae data where we simulated the model only for a time period proportional to the half life of the corresponding transcript [142]. In this case, steady state was not achieved and the translation rate was taken as the mean translation rate over the elapsing time period. This modification however, was unable to improve the predictive power of the model and in effect resulted in an opposite outcome.

9.5.6 Zhang model

Zhang model [114] similar to the ASEP model with the only change that the codon translation times are deterministic.

9.5.7 The relation between translation rate and protein abundance

Here we would like to discuss the relation between translation rates and protein concentration/abundance. In what follows we will provide justification for the intuitive expectation that protein abundance should stand in high positive correlation with translation rates. Generally speaking, protein abundance levels are determined by a balance between protein production and degradation rates. Fixing the degradation rate, protein abundance levels will rise when the production rate is increased. Fixing the production rate, protein abundance levels will decrease when the degradation rate is increased. That said, one must also bear in mind that protein degradation rates are unavailable in most of the analyzed cases. And so, any current real data analysis is forced to average out the effect of protein degradation and focus on the contribution of the production rate to the determination of protein abundance levels.

Let c_i denote the concentration of protein i and let us assume that this protein is translated from a certain mRNA transcript whose copy numbers are

denoted by m_i. In general, the dynamics of this process may be described by the following differential equation: $\frac{dc_i(t)}{dt} = R_i \cdot m_i - D_i(c_i)$. Here R_i and $D_i(c_i)$ are the translation rate per mRNA molecule and the degradation rate of protein i respectively. One possible choice for $D_i(c_i)$ is: $D_i(c_i) = d_i \cdot c_i(t)$ where $d_i > 0$ is constant. Although this is a common approximation we will not base our conclusions on this particular choice and would only require that $D_i(c_i)$ is a monotonically increasing function of the concentration c_i. In general, the function D_i depends on the protein i, i.e., it can be different from protein to protein. Here however, we will replace the protein specific function D_i with a genomic average degradation function D which will be assumed monotonically increasing. Note that by definition, this function does not depend on the index i.

The steady state solution of the above differential equation (with D_i replaced by D) is: $D_i(c_i^{ss}) = R_i \cdot m_i$ where c_i^{ss} is the steady state concentration of the protein i. From the monotonicity of $D(c_i)$ it follows that c_i^{ss} is a monotonically increasing function of $R_i \cdot m_i$. This fact provides justification for the use of $R_i \cdot m_i$ as a predictor for c_i^{ss}, i.e., one expects $R_i \cdot m_i$ and c_i^{ss} to be positively correlated. Indeed, we have shown that this predictor performs very well, see Subsection 9.5.19. We will now show that R_i itself can also be used as a predictor for c_i^{ss}, the advantage of this predictor is that it is solely based on the coding sequence and no additional information is required for its computation.

The set of mRNA copy numbers $\{m_i\}$ may generally depend on the set of translation rates $\{R_i\}$, for example via the concentration of proteins that are involved in mRNA transcription and regulation. Fortunately, it is known that in endogenous genes translation rates are positively correlated with mRNA levels. Highly expressed genes are under selection to have higher mRNA levels, higher translation rate and higher protein abundance (note that this is not a causal relation; see, for example, [89]). Since mRNA levels are positively correlated with translation rates, higher values of R_i do indeed imply higher values of $R_i \cdot m_i$ and vice versa. Since in hetrogenouse gene expression mRNA copy numbers are usually independent of the mRNA variant of the protein, a similar trend is observed in this case as well. In building a predictor which is solely based on coding sequences, these empirical observation provide justification for using R_i as a predictor for c_i^{ss}. Indeed, as we have demonstrated throughout the paper, this predictor out performs other commonly used predictors.

9.5.8 Data

Protein abundance: protein abundance of S. cerevisiae was downloaded from [98, 99]; protein abundance of different versions (with different codon bias) of GFP library in E. coli were downloaded from [89]; Protein abundance of S. pombe were downloaded from [143] and the Protein abundance E. coli were downloaded from [100].
Profiles of Ribosme density: in S. cerevisiae were downloaded from [97].
Folding energies: of the E. coli GFP library was downloaded from [89].
tRNA copy number: of E. coli, S. cerevisiae, and S. pombe were downloaded

from [103]. tRNA levels in diauxic shift in S. cerevisiae were downloaded from [90].
Coding sequences: Coding sequences of S. cerevisiae, E. coli, and S. pombe were downloaded from [103].
Tissue specific gene expression and tAI in Human: the gene expression was downloaded from [144]; the corresponding tAI were downloaded from [112]. Inferred tissue specific tRNA pool in human liver (the tissue where the correlation between the expression levels and translation rate is the highest) was downloaded from [90, 112] based on [145].
mRNA levels: mRNA levels of E. coli were downloaded from [100]; mRNA levels of S. cerevisiae were downloaded from [128]; mRNA levels of S. pombe were downloaded from [143].

9.5.9 Estimating the tAI based values that were used by the model

Our measure was based on the tAI [109]; as describe below, we adjusted it to our model:

Let n_i be the number of tRNA isoacceptors recognizing codon i. Let $tCGN_{ij}$ be the copy number of the $j-th$ tRNA that recognizes the $i-th$ codon, and let S_{ij} be the selective constraint on the efficiency of the codon-anticodon coupling. We define the absolute adaptiveness, W_i, for each codon i as:

$$W_i = \sum_{j=1}^{n_i} (1 - S_{ij})\, tCGN_{ij}$$

The S_{ij}-values can be organized in a vector (S-vector) as described in [109]; each component in this vector is related to one wobble nucleoside-nucleoside paring: I:U, G:U, G:C, I:C, U:A, I:A, etc.

Sensitivity analysis of the tAI of codons to S_{ij}-values in S.cerevisiae showed that one codon (CGA) is extremely sensitive to these S-values. Increasing/decreasing the S-values by ± 0.5 resulted in a change of up to one order of magnitude (usually much less) in all other codons. In the case of CGA, the change was up to 4000 times higher. The tAI of this codon is relatively low and the model is sensitive to this value. Thus, we replaced the W_i of this codon by the mean tAI of this codons over all possible changes (± 0.5) of S_{ij}-values.

From W_i we obtain p_i, which is the probability that a tRNA will be coupled to the codon

$$p_i = \frac{W_i}{\sum_{j=1}^{61} tCGN_{ij}}$$

The expected time on codon i is $t_i = 1/p_i$. The expected time on a site is the sum of times of all the codons in the site.

9.5.10 Computing the bottleneck

The bottleneck was defined as the slowest window in a gene. The time of a window is the sum of times corresponding to its codons; the size of a window is

15 codons (the results were robust to small changes in the size of the window).

9.5.11 Running times

Figure 49 depicts the running time of our model as a function of λ and site size. As can be seen, when the site size is larger than 10 codons, for all λ the typical running time for a gene is less than 0.1 second.

9.5.12 Real and predicted ribosome density profiles

Measurements of ribosome densities in S. cerevisiae at a resolution of single nucleotides were downloaded from [97]. For comparison to the predictions of various models the profiles were aligned to the beginning of the coding sequences (similarly to the way it was done in [90, 103]). We computed and plotted the mean densities in sites of size 15 codons for each of the profiles (measured and predicted).

9.5.13 DTCO and DPCO — estimating the dependence of genes on codon order in terms of translation rate and protein abundance

To estimate the dependence of the translation rate of genes (at their 'working point') on codon order, DTCO, we performed the following steps:

1. Each mRNA transcript was randomly permuted (i.e., codons were randomly shuffled) 10 times. A library of permuted mRNA transcripts, associated with the original transcript, was thus generated and translation rates were computed for each transcript.
2. We then computed, for each gene separately, the standard deviation (stdev) for the set of rates obtained in stage 1.
3. For each gene, the stdev was normalized by the predicted translation rate of the gene (obtained from the un-permuted mRNA transcript). We call this quantity DTCO and we use it as a measure for the dependence of the translation rate on codon order.

To estimate the dependence of protein abundance on the codon order, DPCO, we performed the following steps:

1. The relation between protein abundance and translation rates seems linear on a log-log scale (Figure 50-52); thus, we inferred a liner regressor of the log of protein abundance from the log of the predicted translation rate.
2. For each gene, and for each permutation, protein abundance was estimated via the regressor in (1). The stdev of the PA distribution associated with each gene (i.e., of the library of permuted transcripts) was then computed.
3. For each gene, the stdev of the predicted protein abundance was normalized by the protein abundance of the original (un-permuted) mRNA.

9.5.14 Finding the 'working point' of a gene

To compute the 'working point's of genes in a certain organism we first found the λ where the correlation between the mean predicted translation rate and protein abundance [99, 100, 143] is maximal. We computed the ratio (in percentages) between the mean genomic translation rate at this point and the mean maximal translation rate (for very large λ); let Q% denote this value (93%, 95%, and 99% in S. cerevisiae, S. pombe, and E. coli respectively). The 'working point' of a gene in a certain organism is the λ where the translation rate of the gene is Q% of its maximal translation rate.

9.5.15 Analysis of the data of Burgess-Brown et al.

For each gene we computed the mean ratio between the synthetic version of the gene and its native version over 41 values of λ (between 0.0002 and 0.0094). The empirical p-value for the Spearman correlation is the probability that a random permutation of the two vectors will give higher correlation. It was computed by performing 100 such permutation and computing the Spearman correlation of each of them.

9.5.16 The statistical test used for comparing the genomic ribosomal densities profile to the predicted profiles

The Wilcoxon rank test that we used is a paired non-parametric test where we compared:

1. The vector of distances between the predictions of our model and the real data (a distance for each point).
2. The vector of distances between the predictions of tAI and the real data.
3. The vector of distances between the predictions of Zhang model and the real data.

We compared (1) to (2) and (1) to (3) and checked the following statistical question: "is there an improvement (in terms of the distance between predicted and real data points) when a more sophisticated model (RFM) is used instead of a less sophisticated one (e.g. the tAI)".

9.5.17 Jackknifing to evaluate the robustness of the inferred optimal size of the chunk

Jackknifing (see, e.g., [146]) was performed as described below. Repeat 100 times:

1. Randomly choose 80% of the genes in S. cerevisiae.
2. Find the chunk size that gives the best correlation with protein abundance.

Report the number of cases (0-100) that we get $C = 25$. The result confidence level was 100 demonstrating a very high confidence.

9.5.18 Supplementary Text 1

Justification for using the tAI and the RFM as an predictor of the co-adaptation between codon bias and tRNA pool.

The tAI and the RFM are based on the genomic tRNA copy number (tGCN; when the expression levels of the tRNAs is unknown) as a surrogate measure for the cellular abundances of tRNAs; it is justified by several observations.

First, in the past, in many organisms, it has been observed that the in vivo concentration of a tRNA bearing a certain anticodon is highly proportional to the number of gene copies coding for this tRNA type. Specifically, in S. cerevisiae a correlation of $r = 0.91$ [148] was reported. In B. subtilis, a correlation of 0.86 between tRNA copy number and tRNA abundance was reported [149]. Similarly, previous papers reported about significant correlation between genomic tRNA copy number and tRNA abundance in E. coli [150, 151]. A related interesting result is the analysis of [152] who measured the translation rate of two glutamate codons: GAA and GAG. They found them to have a threefold difference in translation rate (21.6 and 6.4 codons per second, respectively). Remarkably, the w_i of these codons, which is based on the tRNA pool and affinity of codon-anti-codon coupling and is the basis for the tAI calculation, captures the ratio of translation rate between the two codons. Calculating w_i values for E. coli we found that the ratio between the w_i of GAA and GAG is 3.125 (0.5/0.16) as compared to the 3.34 reported in the experiments (21.4/6.4). This result suggests that there is a direct relation between the adaptation of a codon to the tRNA pool, based on the genomic tRNA copy number, and the time it takes to translate it.

Second, a recent study showed that in S. cerevisiae the promoters of many of the tRNA genes have a low predicted affinity to the nucleosome, suggesting a constitutive expression with little transcriptional regulation capacity [153]. Thus, for fully sequenced genomes, the relative concentrations of the various tRNAs in the cell, and therefore the optimality of the various codons in terms of translation, can be approximated using the respective tRNA gene copy numbers in the genome. In addition, and as we have shown in this chapter, measures that are based on tRNA copy number are highly correlated with protein expression levels (see also [108, 154]).

9.5.19 Supplementary Text 2

Endogenous genes in S. cerevisiae, S. pombe, and E. coli: correlation of the predicted rates with protein abundance given mRNA levels and the correlation of the predicted rate multiplied by the mRNA levels with protein abundance.

E. coli — The correlation with PA given mRNA: $r = 0.33$ ($p = 1.84 \cdot 10^{-11}$) for the RFM vs. $r = 0.29$ ($p = 3.95 \cdot 10^{-9}$; 398 genes with PA & mRNA; see Figure 53) for the tAI; the correlation of (Rate * mRNA) with PA ($r = 0.71$ for the RFM vs. $r = 0.7$ for the tAI, $p < 10^{-16}$ in both cases; see Figure 54);

S. pombe — The correlation with PA given mRNA was also higher for the

RFM, $r = 0.39$ ($p < 10^{-16}$) vs. 0.307 for the tAI ($p < 10^{-16}$) for the tAI (see Figure 55). The correlation of (Rate * mRNA) with PA ($r = 0.7$ for the RFM vs. $r = 0.627$ for the tAI, $p < 10^{-16}$ in both cases; see Figure 56);

S. cerevisiae — The correlation with PA given mRNA: $r = 0.32$ for the RFM vs. 0.35 for the tAI ($p < 10^{-16}$ in both cases; see Figure 57). The correlation of (Rate * mRNA) with PA: 0.58 for the RFM vs. 0.58 for the tAI ($p < 10^{-16}$ in both cases; see Figure 58). The correlation with PA $r = 0.49$ for the RFM vs. 0.57 for the tAI ($p < 10^{-16}$ in both cases).

9.5.20 Supplementary Text 3

The predictions of the tAI and translation efficiency profiles of genes.

The tAI is one of the best known codon bias based predictor of protein abundance (see, for example, [108, 109, 111], $r = 0.65$ between tAI and protein abundance in S. cerevisiae). This measure is the mean co-adaptation of the gene's codon to the tRNA pool of an organism without considering the order of the codons. It may sound a bit surprising that such a simple model gives such good performances. One possible explanation is the fact that initiation rate of genes is relatively low. As we have shown, in this regime there are no interactions between ribosomes and the mean "nominal" velocity [90] is a good approximation of the actual one. An alternative or additional explanation for this phenomenon may be the fact that genes (especially highly expressed genes) tend to have a specific design. It was shown that such genes have a non-decreasing profile of translation rate: they start with a region of slower translation rate (30-50 codons); the translation rate afterwards is higher, relatively constant, and usually proportional to the slower beginning ([90]; Figure 59). Such a profile improves the production rate of proteins (the number of proteins per ribosome per time unit, see [90]) and prevents "traffic jams", thus the tAI becomes a good enough predictor. It is possible that that in this regime the tAI performs well (although it is a non-physical and non-causal predictor) simply since the mean speed is a indeed a good approximation of the production rate in this regime (due to the reasons mentioned above).

9.5.21 Supplementary Text 4

The genomic rate of abortion of ribosomes has power law decay.

Measured ribosome density profile appears linear in a log-log graph. Namely the ribosomal density in the $i-th$ site is $C_1 \cdot x^{-\alpha_{real}}$ where $\alpha_{real} = 0.158$ (Figure 30B); our model also predicts a linear line in a log-log graph but with a smaller slope: the genomic ribosomal density in the $i - th$ site is $C_2 \cdot x^{-\alpha_{RFM}}$ where $\alpha_{RFM} = 0.1$ (Figure 30B). These results may suggest that an additional factor $C_2 \cdot x^{-\alpha_{ABR}}$ with $\alpha_{ABR} = 0.058$ should be added to represent other phenomena such as the genomic rate of abortion of ribosomes.

9.5.22 Supplementary Text 5

The initiation rates used in this study are robust and not over-fitted.

To verify the robustness of the initiation rate use in the analysis of heterologous genes we divided each of the datasets to two parts (each with 50% of the genes). In each case we verified that the same initiation rate optimizes the correlation with PA in both parts.

9.5.23 Supplementary Text 6

Tissue-specific translation rates in Human.

In humans and other mammals the correlations between expression levels and codon bias are relativly low [112, 155]. Thus, in the next stage, we aimed at studying how well the predictions of the RFM correlate with the expression levels in human tissues, comparing it with the correlations obtained using the tAI. The correlations with gene expression of 12,173 genes across 73 tissue are depicted in Figure 32C. As can be seen, some of the correlations with the RFM are more than 8 times higher than the correlations with the tAI the mean improvement in percentages of the RFM with respect to the tAI was 80% (Figure 32C, inset). In addition, some of the correlations between tissue-specifc gene expression and tAI are negative while in the case of the RFM all the correlations are positive. Furthermore, in 71% of the tissues the RFM performed better (Figure 32C). These results demonstrate that our model should be more useful than alternative models for analyzing human gene translation. Thus, it should be helpful for analyzing mutations (see, for example [122]) or SNPs (see, for example, [85, 131]) that cause diseases due to problems in gene translation.

9.5.24 Supplementary Figures

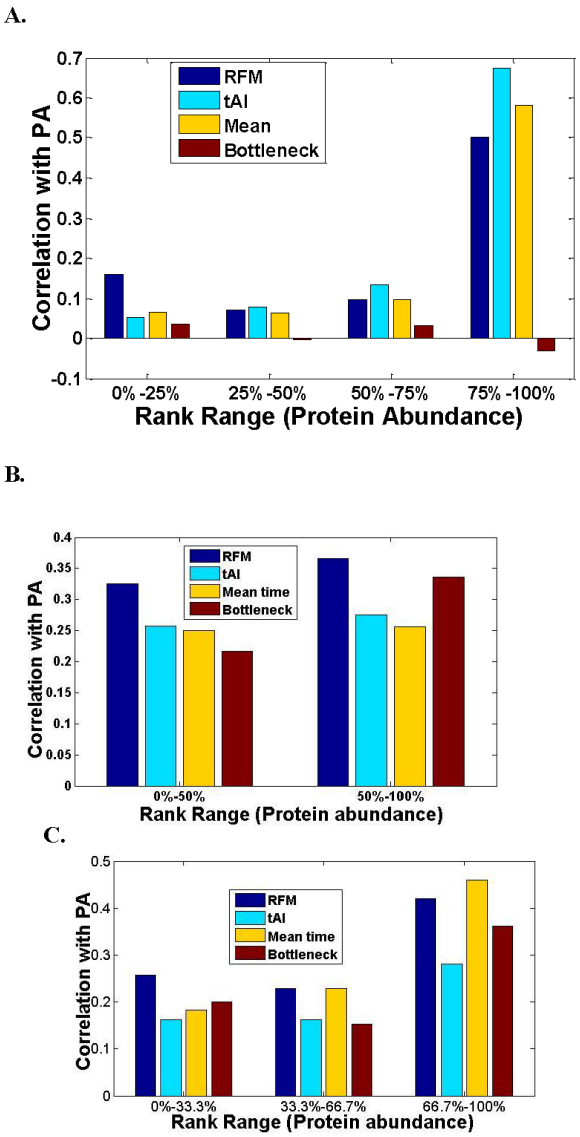

Figure 33: Prediction of protein abundance by the various codon bias based predictors of PA and by the ribosome flow model (RFM) for groups of genes with different levels of protein abundance in S. cerevisiae (A.), E. coli (B.), S. pombe (C.); all bins are of equal size. The RFM outperforms all the other predictors for lowly expressed genes (and in most of the bins) and has significant correlation with PA in all the bins.

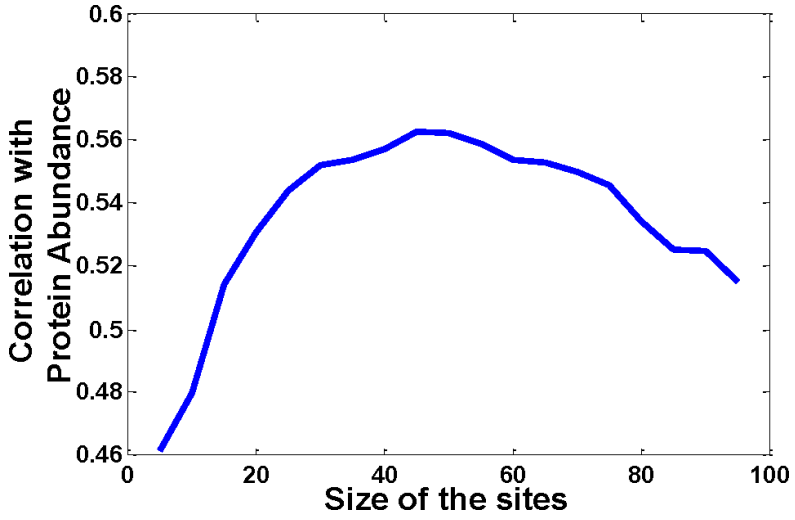

Figure 34: Correlation between protein abundance and the translation rate for various sizes of the translation site unit (C in Figure 1) in E. coli.

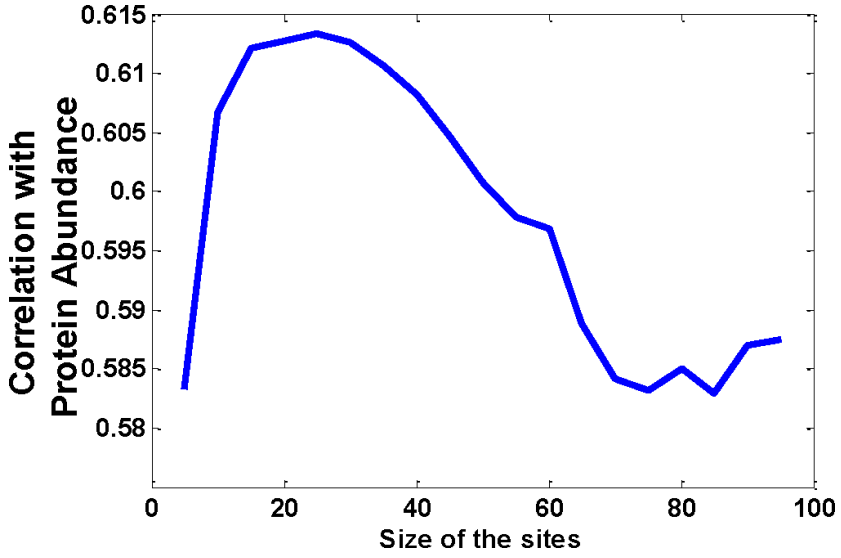

Figure 35: Correlation between protein abundance and the translation rate for various sizes of the translation site unit (C in Figure 1) in S. pombe.

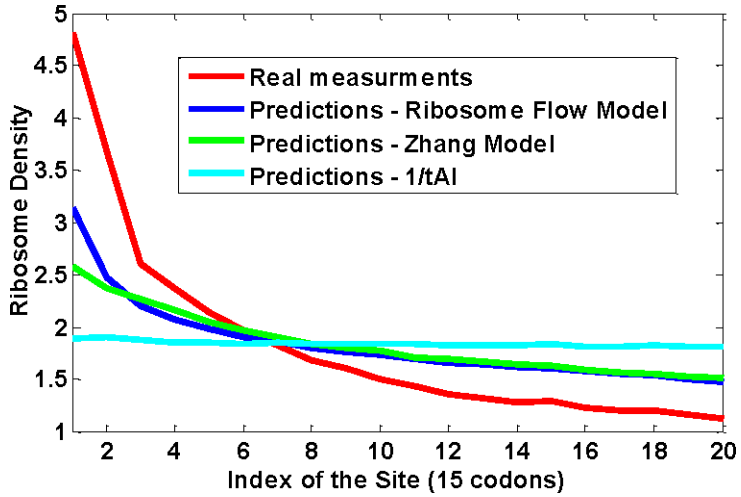

Figure 36: The RFM predicts the genomic profile of ribosome densities in starvation better than the tAI model or the predictor of Zahng et al. All the figures were normalized to have the same mean.

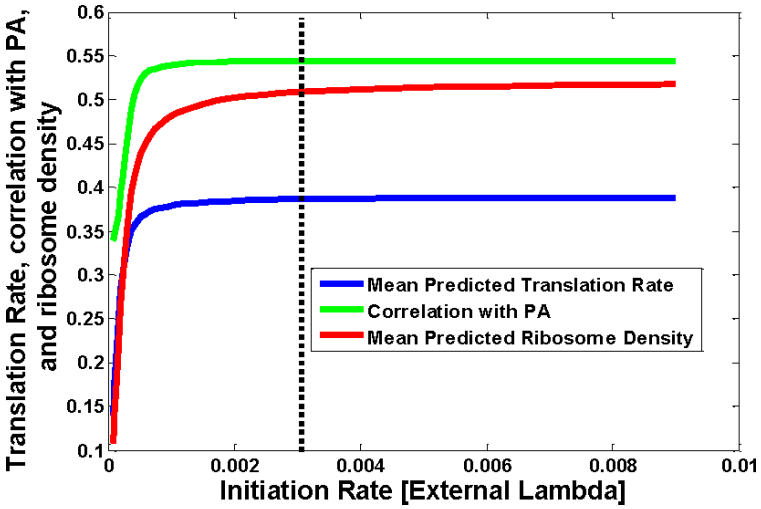

Figure 37: The relation between (the number of available ribosomes in the cell), mean of the translation rate (number of proteins per time unit), and the mean ribosome density in E. coli.

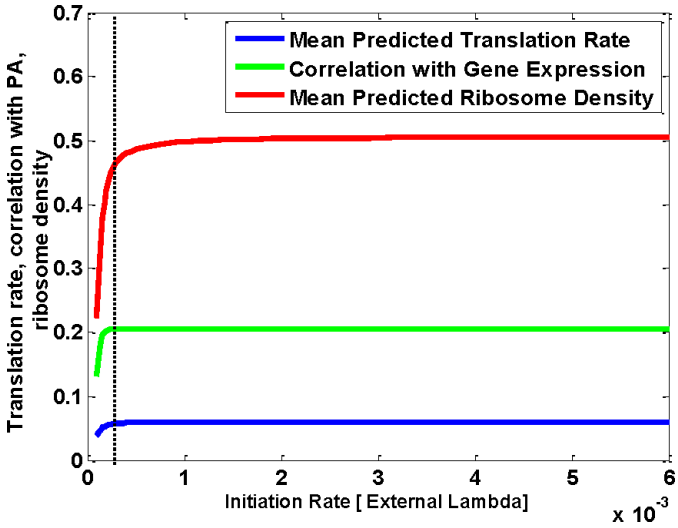

Figure 38: The relation between (the number of available ribosomes in the cell), mean of the translation rate (number of proteins per time unit), and the mean ribosome density in Human liver.

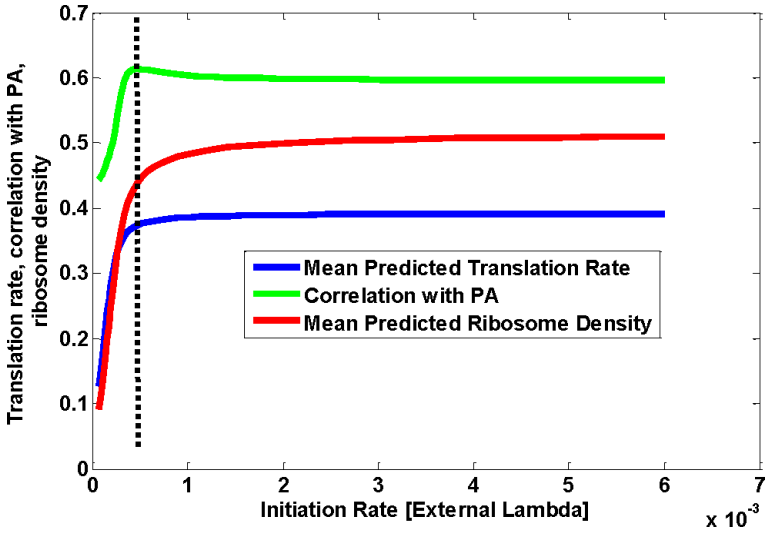

Figure 39: The relation between (the number of available ribosomes in the cell), mean of the translation rate (number of proteins per time unit), and the mean ribosome density in S. pombe.

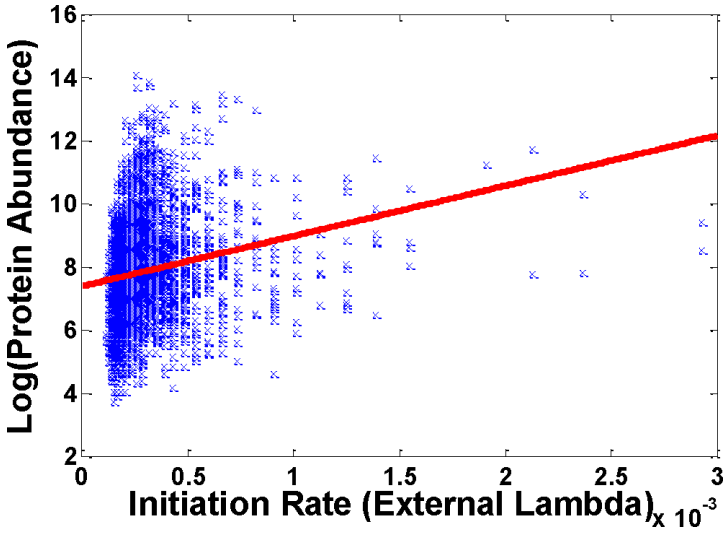

Figure 40: Dot plot — log protein abundance vs. initiation rate in S. cerevisiae.

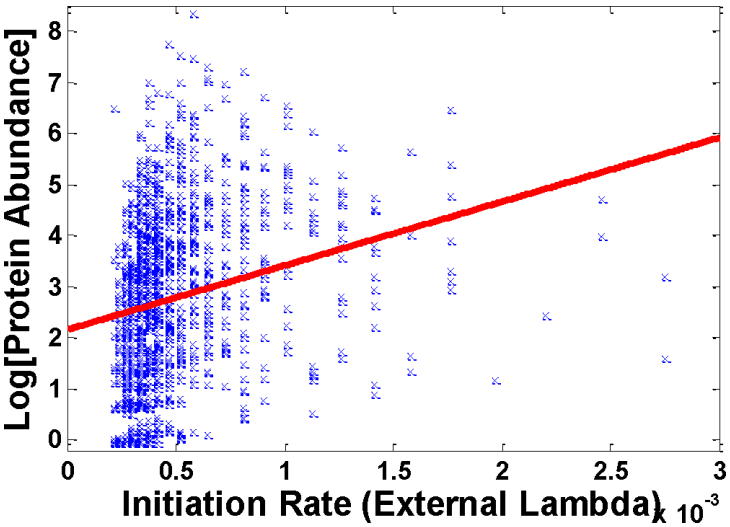

Figure 41: Dot plot — log protein abundance vs. initiation rate in S. pombe.

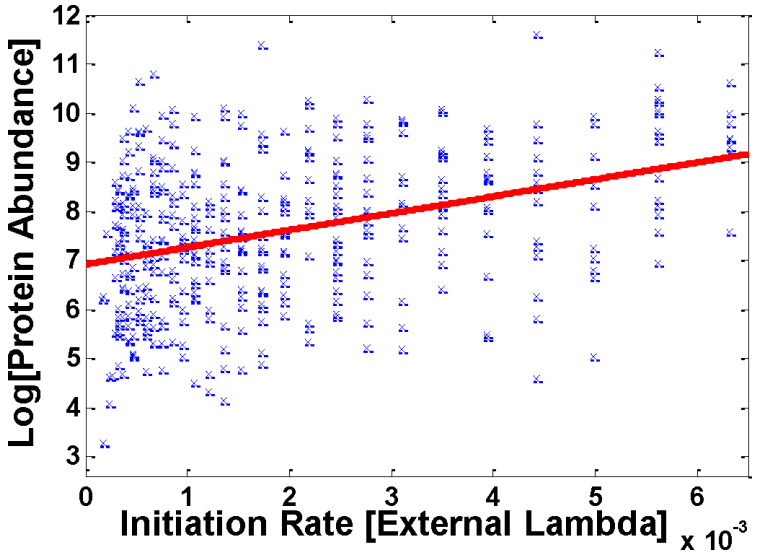

Figure 42: Dot plot — log protein abundance vs. initiation rate in E. coli

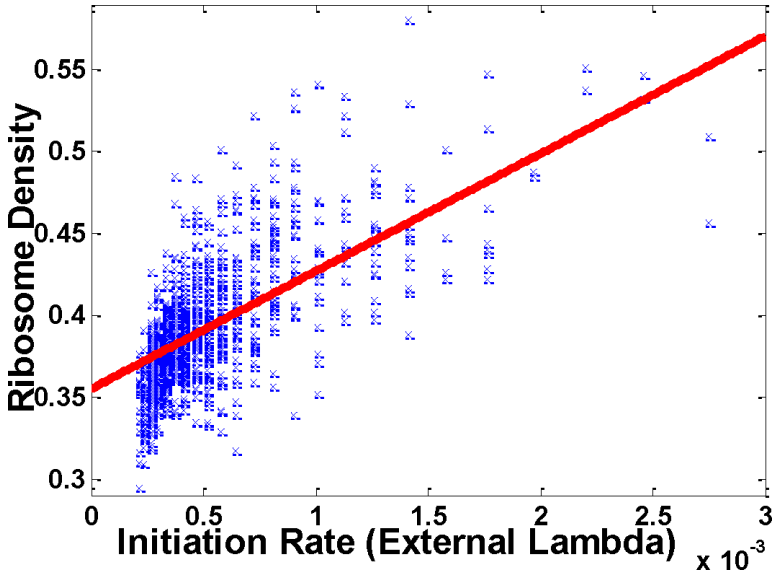

Figure 43: Dot plot — ribosome density vs. initiation rate in S. pombe

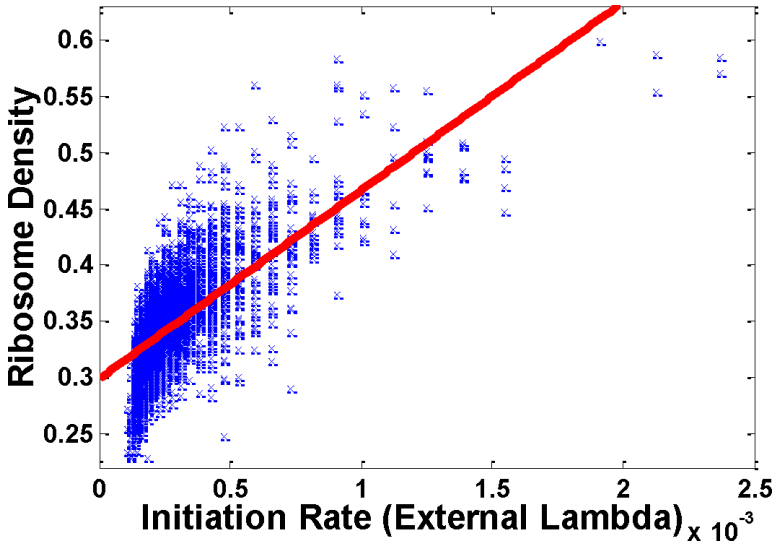

Figure 44: Dot plot — ribosome density vs. initiation rate in S. cerevisiae.

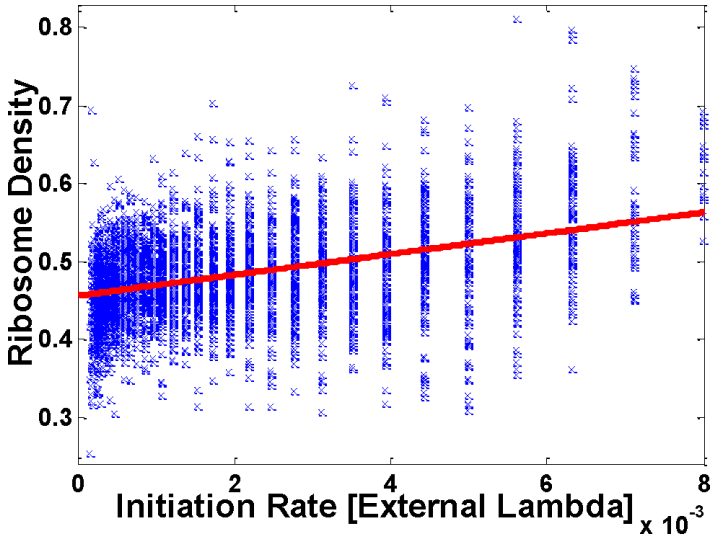

Figure 45: Dot plot — ribosome density vs. initiation rate in E. coli.

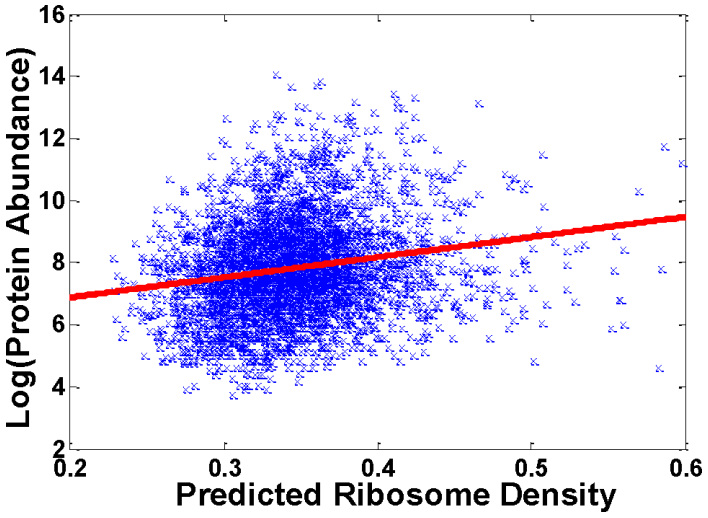

Figure 46: Dot plot — log protein abundance vs. ribosome density in S. cerevisiae.

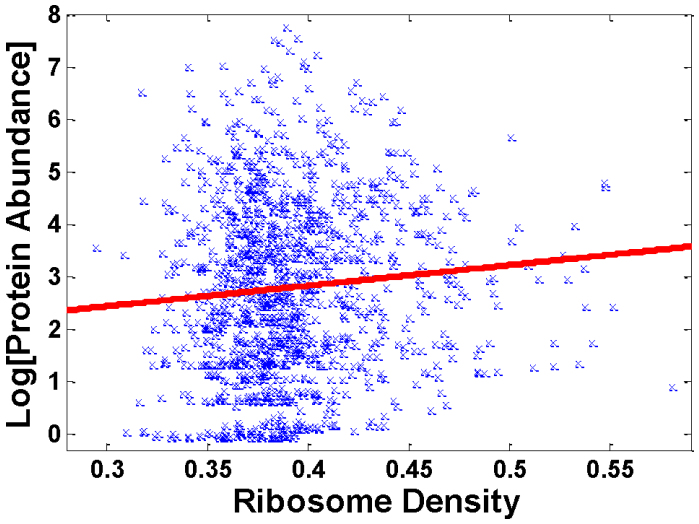

Figure 47: Dot plot — log protein abundance vs. ribosome density in S. pombe.

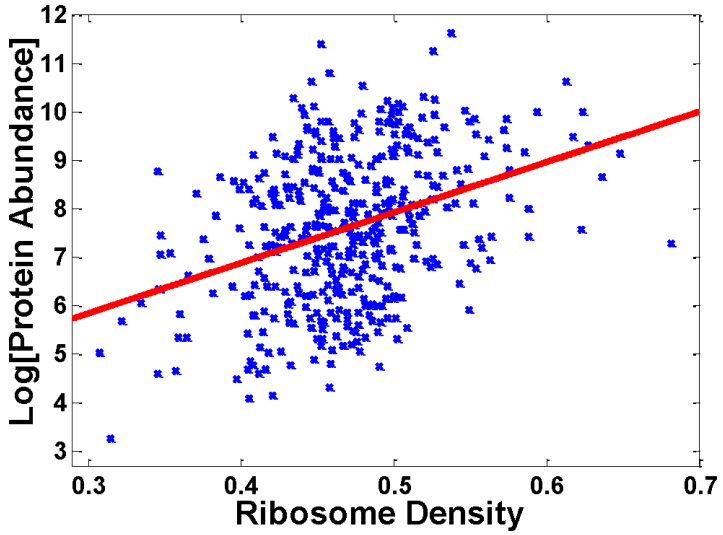

Figure 48: Dot plot — log protein abundance vs. ribosome density in E. coli.

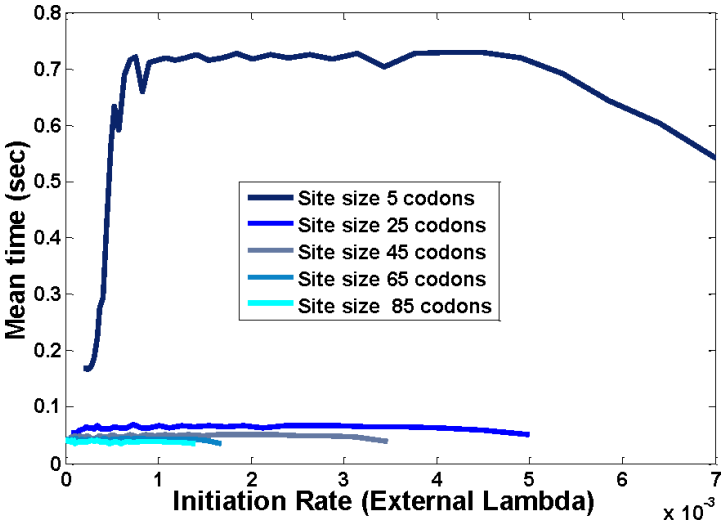

Figure 49: Mean running time (in seconds) for computing the translation rate of the RFM as a function of λ and size of the site.

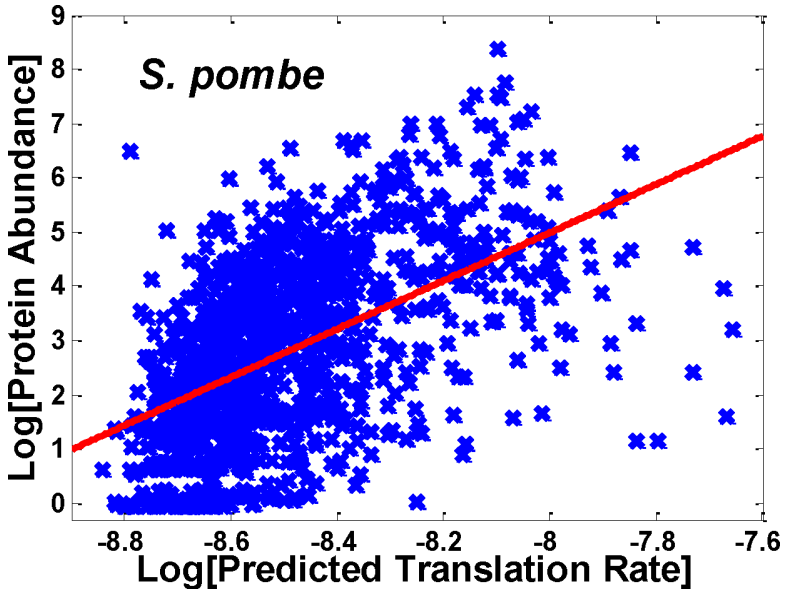

Figure 50: Dot plot – log protein abundance vs. log predicted translation rate in S. pombe.

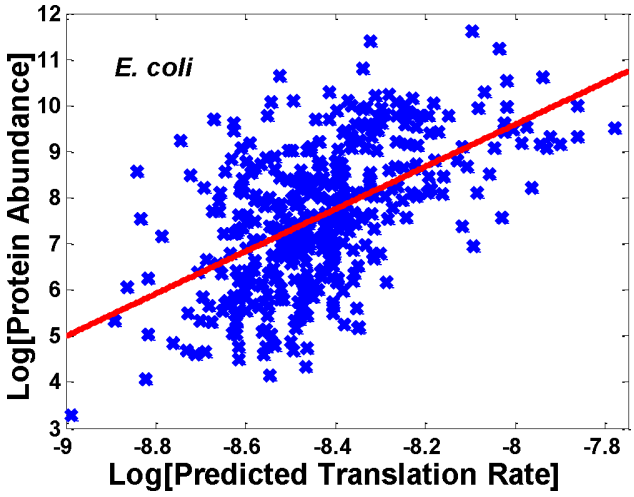

Figure 51: Dot plot – log protein abundance vs. log predicted translation rate in S. pombe.

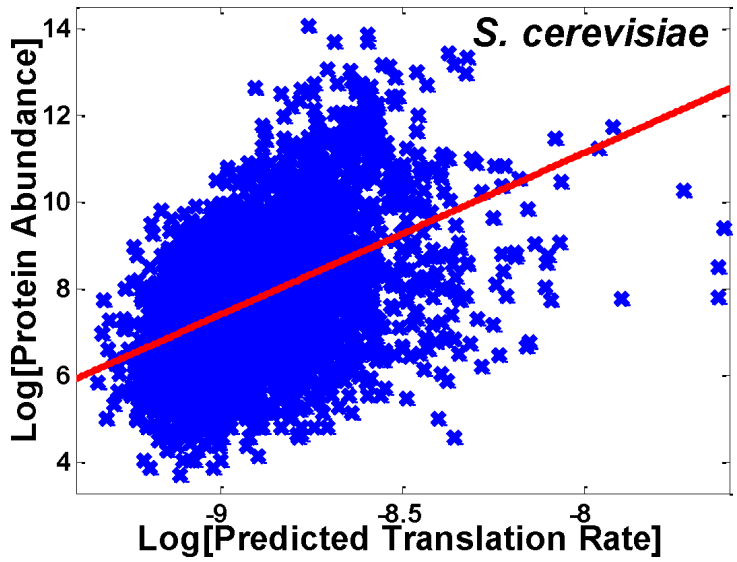

Figure 52: Dot plot – log protein abundance vs. log predicted translation rate in S. cerevisiae.

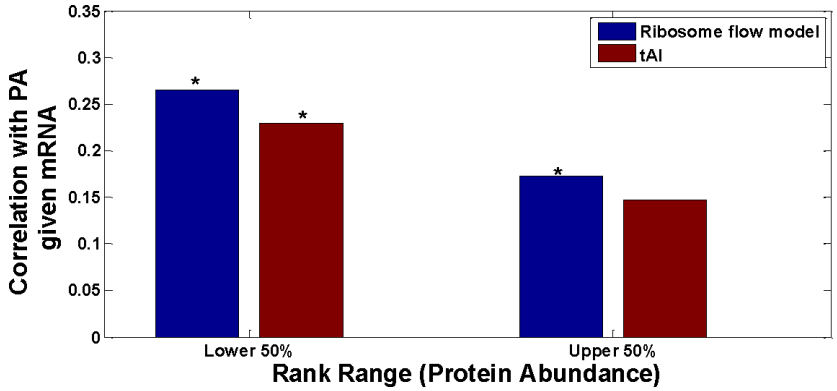

Figure 53: Correlation of the tAI and the RFM and with protein abundance given mRNA levels for groups of genes with different levels of protein in E. coli. All bins are of equal size.

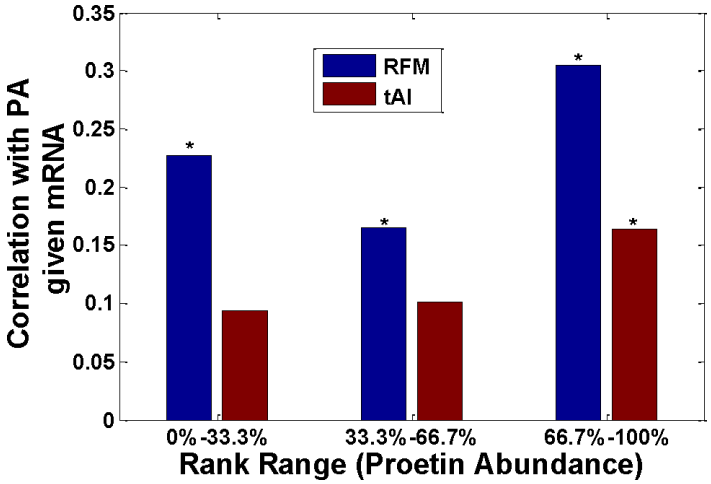

Figure 54: Correlation of the tAI and the RFM with protein abundance given mRNA levels for groups of genes with different levels of protein in S. pombe. All bins are of equal size.

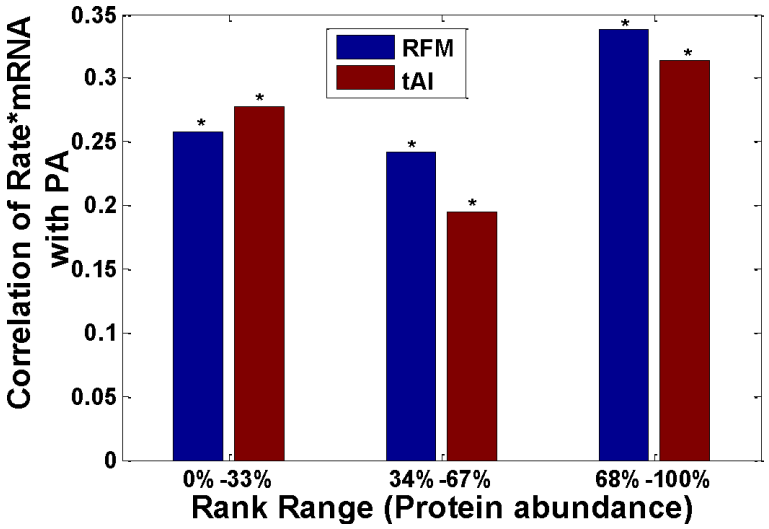

Figure 55: Correlation of the tAI and the RFM with protein abundance multiplied by mRNA levels for groups of genes with different levels of protein in S. pombe. All bins are of equal size.

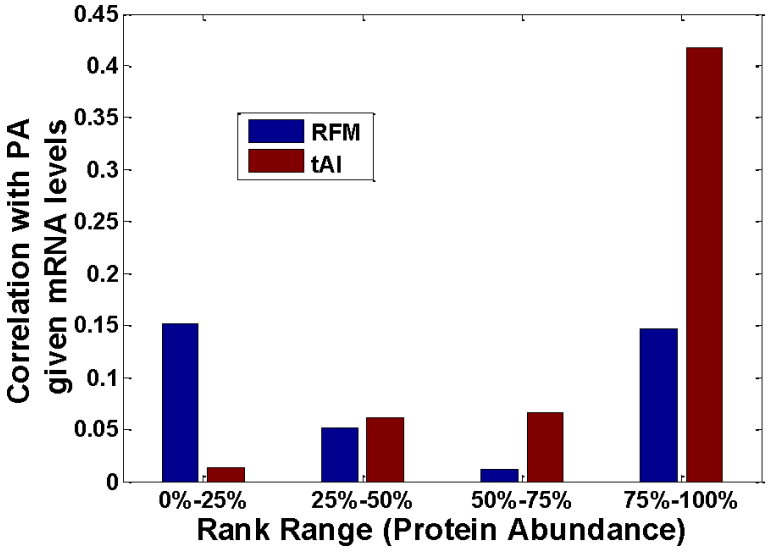

Figure 56: Correlation of the tAI and the RFM with protein abundance given mRNA levels for groups of genes with different levels of protein in S. cerevisiae. All bins are of equal size.

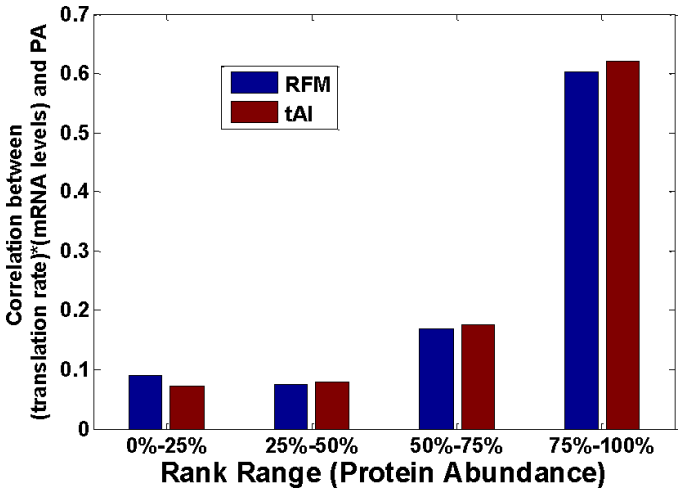

Figure 57: Correlation of the tAI and the RFM with protein abundance multiplies by the mRNA levels for groups of genes with different levels of protein in S. cerevisiae. All bins are of equal size.

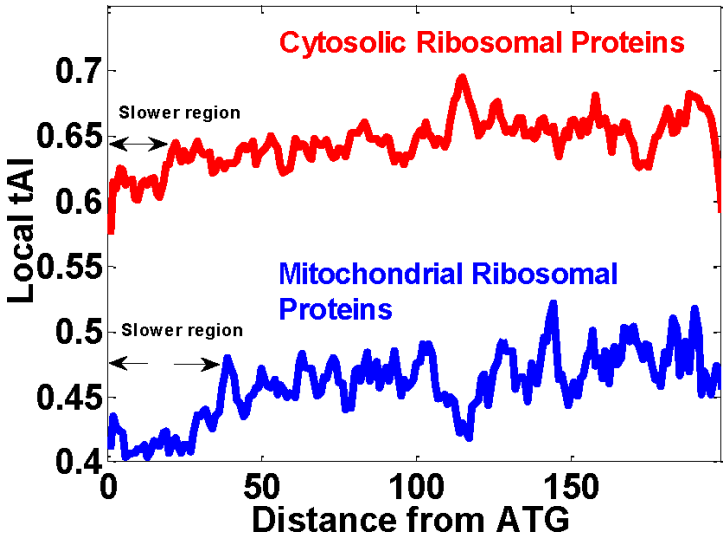

Figure 58: Profiles of tAI of cytosolic and mitochondrial ribosomal proteins in S. cerevisiae.

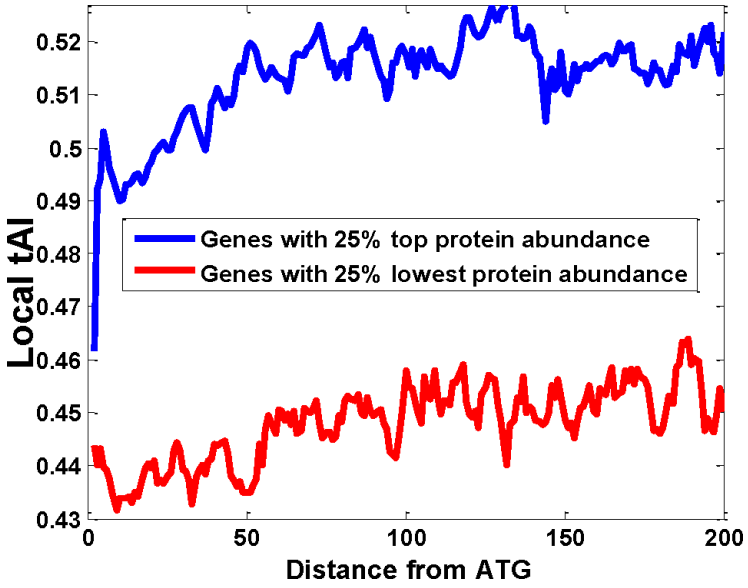

Figure 59: Profiles of tAI of highly expressed genes and lowly expressed genes in S. cerevisiae. Close to the 5' end of the genes there is a region with slower speed. This region is more prominent in highly expressed genes.

References

[1] S. Reuveni, I. Eliazar and U. Yechiali, The Asymmetric Inclusion Process: A Showcase of Complexity. *Phys. Rev. Lett.* **109**, 020603, (2012).

[2] S. Reuveni, I. Eliazar and U. Yechiali, Asymmetric Inclusion Process. *Phys. Rev. E.* **84**, 041101, (2011).

[3] S. Reuveni, I. Eliazar and U. Yechiali, Limit Laws for the Asymmetric Inclusion Process. *Phys. Rev. E.* **86**, 061133, (2012).

[4] S. Reuveni, Catalan's Trapezoids. *To be published in Probability in the Engineering and Informational Sciences.*

[5] S. Reuveni, O. Hirschberg, I. Eliazar and U. Yechiali, Occupation Probabilities and Fluctuations in the Asymmetric Simple Inclusion Process. *Under Review, arXiv:1309.2894.*

[6] S. Reuveni, I. Meilijson, M. Kupiec, E. Ruppin and T. Tuller, Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Mode. *PLoS Computational Biology,* **7**(9), e1002127, (2011).

[7] R. R. P. Jackson, Queueing systems with phase-type service. *Operational Research Quarterly,* **5** (4), 109-120, (1954).

[8] R. R. P. Jackson, Random Queueing Processes with Phase-Type Service. *Journal of the Royal Statistical Society Series B (Methodological)*, **18**, 1, 129-132, (1956).

[9] R. W. Wolff, Stochastic Modeling and the Theory of Queues, Prentice-Hall, (1989).

[10] D. G. Kendall, Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *The Annals of Mathematical Statistics,* **24**, 3, 338, (1953).

[11] J. R. Jackson, Networks of waiting lines. *Operations Research,* **5,** 4, 518-521, (1957).

[12] J. R. Jackson, Jobshop-like Queuing Systems, *Management Science,* **10,** 1, 131-142, (1963).

[13] R. R. P. Jackson, Book review: Queueing networks and product forms: a systems approach. *IMA Journal of Management Mathematics,* **6**, (4), 382-384, (1995).

[14] H. Chen and D. D. Yao, *Fundamentals of Queueing Networks,* Springer, Berlin, (2001).

[15] J. R. Jackson, Comments on "Jobshop-Like Queueing Systems: The Background". *Management Science,* **50,** (12), 1796–1802, (2004).

[16] J. R. Jackson, Jobshop-Like Queueing Systems, *Management Science,* **50**, 12, 1796–1802, (2004).

[17] F. Spitzer, Interaction of Markov processes. *Adv. Math.,* **5**, 246, (1970).

[18] E. Levine, D. Mukamel and G.M. Schutz, Zero-Range Process with Open Boundaries. *Journal of Statistical Physics,* **120**, Nos. 5/6, (2005).

[19] M. R. Evans and T. Hanney, Nonequilibrium statistical mechanics of the zero-range process and related models. *J. Phys. A* Math. *Gen.*, **38**, R195–R240, (2005).

[20] B. Derrida, E. Domany and D. Mukamel. An exact solution of the one dimensional asymmetric exclusion model with open boundaries. *Journal of Statistical Physics,* **69**, 667, (1992).

[21] O. Golinelli and K. Mallick, The asymmetric simple exclusion process: an integrable model for non-equilibrium statistical mechanics. *J. Phys. A*: Math. Gen. **39**, 12679, (2006).

[22] B. Derrida, Non-equilibrium steady states: fluctuations and large deviations of the density and of the current. *J. Stat. Mech.,* P07023, (2007).

[23] R. A. Blythe and M. R., Evans, Nonequilibrium steady states of matrix-product form: a solver's guide. *J. Phys. A: Math. Theor.,* **40**, R333-R441, (2007).

[24] C. T. MacDonald, J. H. Gibbs and A. C. Pipkin, Kinetics of biopolymerization on nucleic acid templates. *Biopolymers,* **6**, 1, (1968).

[25] K. Heckmann, Single file diffusion Passive Permeability of Cell Membranes. *Biomembranes,* **3**, 127, (1972).

[26] D. G. Levitt, Dynamics of a single-file pore: Non-Fickian behavior. *Phys. Rev. A,* **8**, 3050, (1973).

[27] P. M. Richards, Theory of one-dimensional hopping conductivity and diffusion. *Phys. Rev. B,* **16**, 1393, (1977).

[28] B. Widom, J. L. Viovy and A. D. Defontaines, Repton model of gel electrophoresis and diffusion. *J. Physique I,* **1**, 1759, (1991).

[29] M. Schreckenberg and D. E. Wolf (ed), Traffic and Granular Flow, New York: Springer, (1998).

[30] L. B. Shaw, R.K. Zia and K.H. Lee, Totally asymmetric exclusion process with extended objects: a model for protein synthesis. *Phys Rev E,* **68**, 021910, (2003).

[31] T. Halpin-Healy and Y. C. Zhang, Kinetic roughening phenomena, stochastic growth, directed polymers and all that. *Phys. Rep.,* **254**, 215, (1995).

[32] J. Krug, Origins of scale invariance in growth processes. *Adv. Phys.,* **46**, 139, (1997).

[33] R. Bundschuh, Asymmetric exclusion process and extremal statistics of random sequences, *Phys. Rev. E,* **65**, 031911, (2002).

[34] S. Klumpp and R. Lipowsky, Traffic of molecular motors through tube-like compartments, *J. Stat. Phys.,* **113**, 233, (2003).

[35] S. F. Burlatsky, G. S. Oshanin, A. V. Mogutov and M. Moreau, Directed walk in a one-dimensional lattice gas. *Physics Letters A,* **166**, 230-234, (1992).

[36] S. F. Burlatsky, G. Oshanin, M. Moreau and W. P. Reinhardt, Motion of a driven tracer particle in a one-dimensional symmetric lattice gas. *Phys Rev E,* **54**, 3165-3172, (1996).

[37] O. Bénichou, A. M. Cazabat, J. De Coninck, M. Moreau and G. Oshanin, Stokes Formula and Density Perturbances for Driven Tracer Diffusion in an Adsorbed Monolayer. *Phys. Rev. Lett.,* **84**, 511-514, (2000).

[38] C. M. Monasterio and Gleb Oshanin, Bias and bath mediated pairing of particles driven through a quiescent medium. *Soft Matter,* **7**, 993-1000, (2011).

[39] B. Derrida, M.R. Evans, V. Hakim and V. Pasquier, Exact solution of a 1d asymmetric exclusion model using a matrix formulation. *J. Phys. A,* **26**, 1493-1517, (1993).

[40] M. F. Neuts, The Busy Period of a Queue with Batch Service. *Operations Research,* **13**, 815-819, (1965).

[41] H. Kaspi, O. Kella and D. Perry, Dam processes with state dependent batch sizes and intermittent production processes with state dependent rates. *Queueing Systems: Theory and Applications,* **24**, 37-57, (1997).

[42] O. Boxma, D. Perry, W. Stadje and S. Zacks, A Markovian growth-collapse model. *Advances in Applied Probabbility,* **38**, 221-243, (2006).

[43] O. Kella, On growth collapse processes with stationary structure and their shot-noise counterparts. *Journal of Applied Probability,* **46**, 363-371, (2009).

[44] B. J. Martin, Batch queues, reversibility and first-passage percolation. *Queueing Systems: Theory and Applications*, **62**, 411-427, (2009).

[45] P. Bak, How nature works: the science of self organized criticality. *Copernicus*, (1996).

[46] M. G. Rozman, M. Urbach, J. Klafter and F. J. Elmer, Atomic scale friction and different phases of motion of embedded molecular systems. *J. Phys. Chem. B,* **102**, 7924-7930, (1998).

[47] J. M. Carlson, J. S. Langer and B. E. Shaw, Dynamics of earthquake faults. *Rev. Mod. Phys.,* **66**, 657-670, (1994).

[48] I. Eliazar and J. Klafter, A growth-collapse model: Lévy inflow, geometric crashes, and generalized Ornstein-Uhlenbeck dynamics. *Physica A,* **334**, 1-21, (2004).

[49] I. Eliazar and J. Klafter, Stochastic Ornstein-Uhlenbeck capacitors. *Journal of Statistical Physics,* **118**, 177-198, (2005).

[50] I. Eliazar and J. Klafter, Growth-collapse and decay-surge evolutions, and geometric Langevin equations. *Physica A,* **367**, 106-128, (2006).

[51] M. Smoluchowski, Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Lösungen. *Z. phys. Chem.,* **92**, (1917).

[52] I. M. Sokolov, S. B. Yuste, J. J. Ruiz-Lorenzo and K. Lindenberg. Mean field model of coagulation and annihilation reactions in a medium of quenched traps: Subdiffusion. *Phys. Rev. E,* **80**, 051114, (2009).

[53] S. B. Yuste, J. J. Ruiz-Lorenzo and K. Lindenberg. Coagulation reactions in low dimensions: Revisiting subdiffusive A+A reactions in one dimension. *Phys. Rev. E,* **80**, 051114, (2009).

[54] P. L. Krapivsky, S. Redner and E. Ben-Naim. A kinetic view of statistical physics. Cambridge University Press, Cambridge, UK, 2010.

[55] D. Ben-Avraham. The coalescence process, $A + A \to A$, and the method of interparticle distribution functions. In V. Privman, editor, Nonequilibrium Statistical Mechanics in One Dimension, pages 29–50. Cambridge University Press, Cambridge, UK, 2005.

[56] K. Jain and M. Barma, Phases of a conserved model of aggregation with fragmentation at fixed sites. *Phys. Rev. E.,* **64**, 016107, (2001).

[57] Note that $\langle X_k \rangle = \langle X_k \mid X_k > 0 \rangle \Pr(X_k > 0)$, and since $\langle X_k \rangle = 1$ it follows that $\langle X_k \mid X_k > 0 \rangle = 1/\Pr(X_k > 0)$.

[58] S. Redner, *A Guide to First-Passage Processes,* Cambridge University Press, (2001).

[59] W. Feller, An Introduction to Probability Theory and Its Applications, Wiley, second edition, (1991).

[60] S.M. Ross, Applied Probability Models with Optimization Applications, Dover, (1992).

[61] B.B. Mandelbrot, Fractals and scaling in finance, Springer, (1997).

[62] I. Eliazar and J. Klafter, The maximal process of nonlinear shot noise. *Physica A,* **388**, 1755-1779, (2009).

[63] R. G. Bitran and D. Tirupati, Tradeoff Curves, Targeting and Balancing in Manufacturing Queuing Networks. *Operations Research,* **37**, 4, 547-564, (1989).

[64] H. M. Markowitz, Portfolio Selection, *J. Finance,* **7**, 77, (1952).

[65] D. Bertsekas, Convex Analysis and Optimization, Athena Scientific, (2003).

[66] P. Greulicha and A. Schadschneidera, Phase diagram and edge effects in the ASEP with bottlenecks. *Physica A,* **387**, 1972-1986, (2008).

[67] A. B. Kolomeisky, Asymmetric simple exclusion model with local inhomogeneity. *J. Phys. A Math. Gen.,* **31**, 1153, (1998).

[68] I. Eliazar and J. Klafter, On the active periods of nonlinear Shot Noise. *Physica A,* **363**, 237, (2006).

[69] W. M. Louis, A Note on the Busy Period of an *M/G/1* Finite Queue. *Operations Research*, **23**, 6, 1179, (1975).

[70] O. J. Boxma and V. Dumas, The busy period in the fluid queue. *SIGMETRICS Perform. Eval. Rev.,* **26**, 100, (1998).

[71] D. P. Heyman and M. J. Sobel, Stochastic Models in Operations Research, McGraw-Hill, (1982).

[72] H. J. Tijms, Stochastic Modeling and Analysis, John Wiley & Sons, (1986).

[73] P. Billingsley, Probability and Measure, 3rd Edition, John Wiley & Sons, (1995).

[74] A. W. Van der Vaart, Asymptotic statistics, Cambridge University Press, (1998).

[75] K. Thomas, Catalan Numbers with Applications, Oxford University Press, (2008).

[76] H.G. Forder, Some Problems in Combinatorics. *Math. Gaz.,* **45**, 199, (1961).

[77] L. W. Shapiro, A Catalan Triangle. *Disc. Math.,* **14**, 83, (1976).

[78] D. F. Bailey, Counting Arrangements of 1's and -1's. *Math. Mag.,* **69**, 128, (1996).

[79] D.D. Frey and J.A. Sellers, Generalization of the Catalan Numbers. *Fibonacci Quarterly*, **39**, 142-148, (2001).

[80] E. Duchi and G. Schaeffer, A combinatorial approach to jumping particles, *J. Comb. Theory A*, **110**, 1-29, (2005).

[81] G. I. Barenblatt, Scaling, self-similarity, and intermediate asymptotics, Cambridge, Cambridge University Press, (1996).

[82] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, New York, Dover Publications, (1972).

[83] S. Redner, A Guide to First-Passage Processes, Cambridge University Press, (2001).

[84] S. Uemura, C.E. Aitken, J. Korlach, B.A. Flusberg and S.W. Turner et al., Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature,* **464**, 1012-1017, (2010).

[85] C. Kimchi-Sarfaty, J.M. Oh, I.W. Kim, Z.E. Sauna, A.M. Calcagno, et al., A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science,* **315**, 525-528, (2007).

[86] I. Bahir, M. Fromer and Y. Prat, M. Linial, Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol,* **5**, 311, (2009).

[87] D.A. Drummond and C.O. Wilke, Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, **134**, 2, 341-352, (2008).

[88] C. Gustafsson, S. Govindarajan and J. Minshull, Codon bias and heterologous protein expression. *Trends Biotechnol,* **22**, 346-353, (2004).

[89] G. Kudla, A.W. Murray, D. Tollervey and J.B. Plotkin, Coding-sequence determinants of gene expression in Escherichia coli., *Science,* **324**, 255-258, (2009).

[90] T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, et al., An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell,* **141**, 344-354, (2010).

[91] S.C. Wenzel and R. Müller, Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways. *Curr Opin Biotechnol,* **16**, 594-606, (2005).

[92] K.B. Scholten, D. Kramer, E.W. Kueter, M. Graf, T. Schoedl, et al., Codon modification of T cell receptors allows enhanced functional expression in transgenic human T cells. *Clin Immunol,* **119**, 135-145, (2006).

[93] Y. Arava, Y. Wang, J.D. Storey, C.L. Liu, P.O. Brown, et al., Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proc Natl Acad Sci,* **100**, 3889-3894, (2003).

[94] T. Warnecke and L.D. Hurst, GroEL dependency affects codon usage support for a critical role of misfolding in gene evolution. *Mol Syst Biol,* **6**, 340, (2010).

[95] J.A. van den Berg, K.J. van der Laken, A.J. van Ooyen, T.C. Renniers, K. Rietveld, et al., Kluyveromyces as a host for heterologous gene expression: expression and secretion of prochymosin. *Biotechnology,* **8**, 135-139, (1990).

[96] G. Lithwick and H. Margalit, Relative predicted protein levels of functionally associated proteins are conserved across organisms. *Nucleic Acids Res,* **33**, 1051-1057, (2005).

[97] N.T. Ingolia, S. Ghaemmaghami, J.R. Newman and J.S. Weissman, Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science,* **324**, 218-223, (2009).

[98] J.R. Newman, S. Ghaemmaghami, J. Ihmels, D.K. Breslow, M. Noble, et al., Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature,* **441**, 840-846, (2006).

[99] S. Ghaemmaghami, W.K. Huh, K. Bower, R.W. Howson, A. Belle, et al., Global analysis of protein expression in yeast. *Nature,* **425**, 737-741, (2003).

[100] P. Lu, C. Vogel, R. Wang, X. Yao and E.M. Marcotte, Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol,* **25**, 117-124, (2007).

[101] K.A. Dittmar, E.M. Mobley, A.J. Radek and T. Pan, Exploring the regulation of tRNA distribution on the genomic scale. *J Mol Biol,* **337**, 31-47, (2004).

[102] Y. Taniguchi, P.J. Choi, G.W. Li, H. Chen, M. Babu, et al., Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533-538, (2010).

[103] T. Tuller, Y.Y. Waldman, M. Kupiec and E. Ruppin, Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci*, **107**, 3645-3650, (2010).

[104] M. Welch, S. Govindarajan, J.E. Ness, A. Villalobos, A. Gurney, et al., Design parameters to control synthetic gene expression in Escherichia coli. *PLoS One,* **4**, e7002, (2009).

[105] K. Fredrick and M. Ibba, How the sequence of a gene can tune its translation. *Cell*, **141**, 227-229, (2010).

[106] G. Cannarozzi, N.N. Schraudolph, M. Faty, P. von Rohr, M.T. Friberg, et al., A role for codon order in translation dynamics. *Cell,* **141**, 355-367, (2010).

[107] R. Heinrich and T.A. Rapoport, Mathematical modeling of translation of mRNA in eucaryotes; steady state, time-dependent processes and application to reticulocytes. *J Theor Biol,* **86**, 279-313, (1980).

[108] T. Tuller, M. Kupiec and E. Ruppin, Determinants of protein abundance and translation efficiency in S. cerevisiae. *PLoS Comput Biol*, **3**, e248, (2007).

[109] M. dos Reis, R. Savva and L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res,* **32**, 5036-5044, (2004).

[110] P.M. Sharp and W.H. Li, The codon Adaptation Index — a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res,* **15**, 1281-1295, (1987).

[111] O. Man and Y. Pilpel, Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet,* **39**, 415-421, (2007).

[112] Y.Y. Waldman, T. Tuller, T. Shlomi, R. Sharan and E. Ruppin, Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Res,* **38**, 2964-2974, (2010).

[113] G. Zhang and Z. Ignatova, Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PLoS One,* **4**, e5036, (2009).

[114] S. Zhang, E. Goldman and G. Zubay, Clustering of low usage codons and ribosome movement. *J Theor Biol*, **170**, 339-354, (1994).

[115] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, et al., Molecular Biology of the Cell, New York, (2002).

[116] M. Kaczanowska and M. Ryden-Aulin, Ribosome biogenesis and the translation process in Escherichia coli., *Microbiol Mol Biol Rev,* **71**, 477-494, (2007).

[117] D. Zenklusen, D.R. Larson and R.H. Singer, Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol,* **15**, 1263-1271, (2008).

[118] J.R. Warner, The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci*, **24**, 437-440, (1999).

[119] R. Kawaguchi and J. Bailey-Serres, mRNA sequence features that contribute to translational regulation in Arabidopsis. *Nucleic Acids Res,* **33**, 955-965, (2005).

[120] H. Miyasaka, The positive relationship between codon usage bias and translation initiation AUG context in Saccharomyces cerevisiae. *Yeast,* **15**, 633-637, (1999).

[121] Y. Osada, R. Saito, M. Tomita, Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes. *Bioinformatics,* **15**, 578-581, (1999).

[122] Y.Y. Waldman, T. Tuller, R. Sharan and E. Ruppin, TP53 cancerous mutations exhibit selection for translation efficiency. *Cancer Res*, **69**, 8807-8813, (2009).

[123] N.R. Voss, M. Gerstein, T.A. Steitz and P.B. Moore, The geometry of the ribosomal polypeptide exit tunnel. *J Mol Biol,* **360**, 893-906, (2006).

[124] G. Zhang, M. Hubalewska and Z. Ignatova, Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol,* **16**, 274-280, (2009).

[125] K. Fredrick and M. Ibba, How the sequence of a gene can tune its translation. *Cell,* **141**, 227-229, (2009).

[126] N. Ban, P. Nissen, J. Hansen, P.B. Moore and T.A. Steitz, The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science,* **289**, 905-920, (2000).

[127] A. Basu and D. Chowdhury, Traffic of interacting ribosomes: effects of single-machine mechanochemistry on protein synthesis. *Phys Rev E,* **75**, 021902, (2007).

[128] Y. Wang, C.L. Liu, J.D. Storey, R.J. Tibshirani, D. Herschlag, et al. Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci,* **99**, 5860-5865, (2002).

[129] N.A. Burgess-Brown, S. Sharma, F. Sobott, C. Loenarz, U. Oppermann, et al. Codon optimization can improve expression of human genes in Escherichia coli: A multi-gene study. *Protein Expr Purif*, **59**, 94-102, (2008).

[130] J.L. DeRisi, V.R. Iyer and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* **278**, 680-686, (1997).

[131] J.M. Comeron, Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. *Genetics,* **167**, 1293-1304, (2004).

[132] J.M. Comeron, Weak selection and recent mutational changes influence polymorphic synonymous mutations in humans. *Proc Natl Acad Sci,* **103**, 6940-6945, (2006).

[133] C. Vogel, S. Abreu Rde, D. Ko, S.Y. Le, B.A. Shapiro, et al., Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol,* **6**, 400, (2010).

[134] M. Welch, A. Villalobos, C. Gustafsson and J. Minshull, You're one in a googol: optimizing genes for protein expression. *J R Soc Interface,* **6**, Suppl 4, S467-476, (2009).

[135] G. Wu, Y. Zheng, I. Qureshi, H.T. Zin, T. Beck, et al., SGDB: a database of synthetic genes re-designed for optimizing protein over-expression. *Nucleic Acids Res,* **35**, D76-79, (2007).

[136] G. Wu, L. Dress and S.J. Freeland, Optimal encoding rules for synthetic genes: the need for a community effort. *Mol Syst Biol,* **3**, 134, (2007).

[137] G. Libertini and A. Di Donato, Computer-aided gene design. *Protein Eng,* **5**, 821-825, (1992).

[138] A.K. Sharma and D. Chowdhury, Quality control by a mobile molecular workshop: quality versus quantity. *Phys Rev E,* **82**, 031912, (2010).

[139] A.K. Sharma and D. Chowdhury, Distribution of dwell times of a ribosome: effects of infidelity, kinetic proofreading and ribosome crowding. *Phys Biol,* **8**, 026005, (2011).

[140] A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, et al. Noise in protein expression scales with natural protein abundance. *Nat Genet,* **38**, 636-643, (2006).

[141] P. Pierobon, A. Parmeggiani, F. von Oppen and E. Frey, Dynamic correlation functions and Boltzmann-Langevin approach for driven one-dimensional lattice gas. *Phys Rev E,* **72**, 036123, (2005).

[142] O. Shalem, O. Dahan, M. Levo, M.R. Martinez, I. Furman et al., Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Mol Syst Biol,* **4**, 223, (2008).

[143] M.W. Schmidt, A. Houseman, A.R. Ivanov and D.A. Wolf, Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. *Mol Syst Biol,* **3**, 79, (2007).

[144] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, et al., A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci,* **101**, 6062-6067, (2004).

[145] K.A. Dittmar, J.M. Goodenbour and T. Pan, Tissue-Specific Differences in Human Transfer RNA Expression. *PLoS Genet,* **2**, e221, (2006).

[146] J. Shao and T. Dongsheng, The Jackknife and Bootstrap: Springer-Verlag, Inc., (1995).

[147] L. M. de Godoy, J.V. Olsen, J. Cox, M.L. Nielsen, N.C. Hubner, et al., Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature,* **455**, 1251-1254, (2008).

[148] R. Percudani, A. Pavesi and S. Ottonello, Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J Mol Biol,* **268**, 322-330, (1997).

[149] S. Kanaya, Y. Yamada, Y. Kudo and T. Ikemura, Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**, 143-155, (1999).

[150] T. Ikemura, Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol*, **151**, 389-409, (1981).

[151] H. Dong, L. Nilsson and C.G. Kurland, Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol,* **260**, 649-663, (1996).

[152] M.A. Sorensen and S. Pedersen, Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J Mol Biol,* **222**, 265-280, (1991).

[153] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, et al., A genomic code for nucleosome positioning. *Nature,* **442**, 772-778, (2006).

[154] O. Man and Y. Pilpel, Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat Genet,* **39**, 415-421, (2007).

[155] A.O. Urrutia and L.D. Hurst, Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics,* **159**, 1191-1199, (2001).

[156] J.V. Chamary, J.L. Parmley, L.D. Hurst, Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet,* **7**, 98-108, (2006).