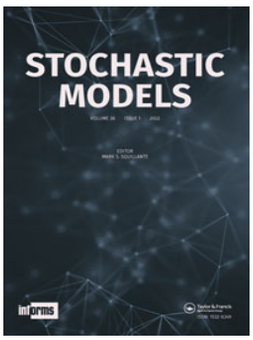

Stochastic Models

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lstm20

Weighted Dyck paths and nonstationary queues

Gianmarco Bet, Jori Selen & Alessandro Zocca

To cite this article: Gianmarco Bet, Jori Selen & Alessandro Zocca (2022): Weighted Dyck paths and nonstationary queues, Stochastic Models, DOI: 10.1080/15326349.2021.2011748

To link to this article: https://doi.org/10.1080/15326349.2021.2011748

Published online: 08 Jan 2022.

STOCHASTIC MODELS
https://doi.org/10.1080/15326349.2021.2011748

Taylor & Francis
Taylor & Francis Group

Weighted Dyck paths and nonstationary queues

Gianmarco Bet[a], Jori Selen[b], and Alessandro Zocca[c]

[a]Mathematics and Computer Science, Università degli Studi di Firenze, Firenze, Italy; [b]ASML, Eindhoven, The Netherlands; [c]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

ABSTRACT

We consider a model for a queue in which only a fixed number N of customers can join. Each customer joins the queue independently at an exponentially distributed time. Assuming further that the service times are independent and follow an exponential distribution, this system can be described as a two-dimensional Markov chain on a finite triangular region $\mathcal{S}$ of the square lattice. We interpret the resulting random walk on $\mathcal{S}$ as a Dyck path that is weighted according to some state-dependent transition probabilities that are constant along one axis, but are rather general otherwise. We untangle the resulting intricate combinatorial structure by introducing appropriate generating functions that exploit the recursive structure of the model. This allows us to derive an explicit expression for the probability mass function of the number of customers served in any busy period (equivalently, of the length of any excursion of the Dyck path above the diagonal) as a weighted sum with alternating sign over a certain subclass of Dyck paths.

ARTICLE HISTORY
Received 19 January 2021
Accepted 24 November 2021

KEYWORDS
Queueing theory; time-inhomogeneous Markov chain; Dyck paths; continuous-time Markov chain

1. Introduction

Time-dependent queueing models are powerful tools for the analysis of real-life situations where the long-term behavior of a system is not a good approximation for its performance. Examples of applications include call centers[6], airline check-in counters[22,30], vaccination hubs[12], outpatient wards of hospitals where the server operates only over a finite amount of time[18,19], and optimal outpatient appointment scheduling[17]. On the other hand, rigorous and explicit results on time-dependent models are mostly out of reach because the standard tools of renewal theory and ergodic theory are often not applicable. In this article we focus on a certain class of time-dependent models called *transitory queueing systems*, introduced by Honnappa and Ward[16], and defined as systems that operate

CONTACT Gianmarco Bet gianmarco.bet@unifi.it Università degli Studi di Firenze, Mathematics and Computer Science, viale Morgagni 67/a, Firenze, 50121 Italy.

only during a finite time horizon. Thus only the time-dependent behavior is of interest. Hence transitory queueing systems are time-dependent models that present even greater technical challenges because their steady-state distribution is trivial (all the probability mass is concentrated in zero). One common approach to tackle this issue is to introduce a scaling parameter N in the queueing model and approximate the resulting system with the asymptotic model obtained by taking $N \to \infty$. This approximation is justified in terms of stochastic process limits, see e.g.,[31,32] and references therein. This approach is robust because it relies on a functional Central Limit Theorem and it has proven to be highly successful. However, this approach has two drawbacks. First, the asymptotic results yield precise approximations only for very large N, and often accurate error estimates are not available. Second, the asymptotic model is often still too complicated to be analyzed exactly, and thus further approximations are needed. In this article we aim at developing novel tools for the analysis of transient queueing systems that do not rely on *any* approximation scheme and that provide *explicit* formulas for the relevant performance metrics. In this article we focus on the number of customers served during a busy period as a proxy for system performance. We emphasize that our approach is not meant to replace the classical asymptotic approximation scheme, but rather to complement it when the approximations it provides are unreliable or analytically intractable.

The canonical model for the study of transitory queueing systems is the so-called $\Delta_{(i)}/G/1$ model[15,16] in which a single queue serves a finite pool of N potential customers, where N will be fixed throughout this article. Each customer joins the queue at a time T_i, where $(T_i)_{i=1}^N$ are positive i.i.d. random variables. Once in the queue, customers are served in a first-come-first-served fashion. Each customer requires an amount of service S_i, where are i.i.d. random variables which are independent from the T_i. Once a customer is served, they leave the system permanently. The $\Delta_{(i)}/G/1$ model was first introduced in Honnappa and Jain[14], where it emerged as the solution of a game-theoretic optimization problem in a queueing setting. Furthermore, in Honnappa and Ward[16] it was proven that, under the appropriate scaling, several other transitory models have the same asymptotic behavior as the $\Delta_{(i)}/G/1$ model. Hence, the $\Delta_{(i)}/G/1$ model should be seen as the canonical transitory queueing model, similarly as how the $G/G/1$ queue is the canonical stationary queueing model. The asymptotic regime $N \to \infty$ of the $\Delta_{(i)}/G/1$ queue has been studied extensively in recent years. In Honnappa et al.[15] the authors prove a functional Law of Large Numbers (fLLN) and a functional Central Limit Theorem (fCLT) for the queue-length process. They identify the limit processes explicitly, but these are considerably difficult to analyze and explicit formulas for

quantities of interest are not available. In a series of works[2–5] the authors consider the $\Delta_{(i)}/G/1$ queue in the heavy-traffic regime that is obtained by assuming the instant of peak congestion is at $t=0$. Their results are also fCLT's for the queue-length process. In all the cases, the limit process is a reflected stochastic process with negative quadratic drift, for which several explicit expressions for quantities of interest are available, see Bet et al.[4] for details.

Here we offer a new perspective on the $\Delta_{(i)}/G/1$ model, which we now summarize. We assume that the arrival times T_i are exponentially distributed with rate λ, and that the service times S_i are exponentially distributed with mean $1/\mu$. We focus on the embedded Markov chain associated to the queueing process, and we show that the path of the Markov chain is a Dyck path of order N, that is, a staircase walk in $\mathbb{N}^2$ from (0, 0) to (N, N) that stays above (but may touch) the diagonal. It follows that the transition probabilities of the Markov chain induce a probability measure on the space of Dyck paths. Our result is then an explicit expression for the probability mass function of the excursion lengths of the Dyck path above the diagonal as a weighted sum over a certain subclass of Dyck paths that, roughly speaking, do not fluctuate between any two consecutive diagonal visits. Furthermore, we show that our result holds for general transition probabilities that include the transition probabilities associated with the $\Delta_{(i)}/G/1$ model.

1.1. Related work

1.1.1. Relation to combinatorics and path counting

Dyck paths are some of the most well-studied objects in combinatorics and thus the literature on the subject is vast. Perhaps closest to our approach is the work of Viennot[29]. That article finds general relationships between a certain class of orthogonal polynomials and weighted Motzkin paths, which are a generalization of Dyck paths that allow for diagonal jumps. In particular, Viennot shows that the elements of the inverse coefficient matrix of the polynomials are related to the sum of the weights of all Motzkin paths starting in (0, 0) and with varying length and endpoint. This is in line with our proof technique for Proposition 3.3. A Dyck path may be seen as a certain type of *list structure* where only insertions and deletions are allowed. In Louchard[20], several list structures are considered and it is shown that, if one assumes a uniform distribution over all admissible lists, the resulting stochastic process converges to an appropriate Markov process. This is in contrast with our setting, where the probability of a path depends crucially on the spatial structure of the path. Moreover, we do not obtain asymptotic results, rather exact results that hold for paths of finite length. In Haug and

Prellberg[13], the authors are able to compute the generating function of Dyck paths weighted according to both their area and length. However, their result hinges on asymptotic approximations as the argument of the generating function tends to one. Furthermore Van Rensburg et al.[28], investigates, among others, the enumeration problem for Dyck paths constrained to lie inside a wedge. See the references in Van Rensburg et al.[28] for an overview of the vast literature on the subject. Weighted lattice walks have also received some attention lately. In Courtiel et al.[8] the authors consider weighted paths in the quarter-plane under the assumptions that the weights are central. A weight is central if paths with the same start point, end point and length have the same weight. This symmetry allows for explicit expressions for the generating functions. In our case the start and end points are fixed, but the weight of a path depends not only on its length, and so we cannot exploit the same techniques as Courtiel et al.[8].

1.1.2. Relation to queueing theory

The transient behavior of queueing system is challenging to study and, as such, few results are available on transient behavior of the $\Delta_{(i)}/G/1$ model and related models. Nevertheless, in this section we position our work in the literature and we compare it with various approaches that focus on the steady-state behavior of similar models.

As mentioned in the introduction, our model is instrumental to capture the process taking place at, e.g., an any boarding gate, since, ignoring the no-shows, there is a fixed total number N of passenger that all need to board. Motivated by this application, a model identical to ours was already considered in Wang et al.[30], but the main results therein concern (i) the expected "makespan", that is the total amount of time requested to process/board all passengers (i.e., the expected hitting time of the state (N, N)) and (ii) the probability that the n-th customer finds, upon arrival, i customers already in the system. On the other hand, a slightly different model was considered in Parlar and Sharafali[22]: differently from us, the authors made the simplifying (yet unrealistic) assumption that clerks work proportionally faster when there are more customer queuing. In this setting[22], shows how to derive transient occupancy probabilities by solving a system of differential equations.

The $\Delta_{(i)}/G/1$ model is also intimately connected with queues with Markov-modulated arrival processes, the simplest of which is the MMP/M/1 queue. In a MMP/M/1 queue, customers arrive according to a Poisson process with a *random* intensity that depends on an underlying Markov chain. Therefore, the $\Delta_{(i)}/G/1$ model can be seen as a MMP/M/1 queue, where the underlying Markov chain is a simple pure-death process. There is a rich literature dedicated to the analysis of the MMP/M/1 queue as well

as its numerous generalizations. The early works[25,26] deal with numerical approximations of such systems. In particular[26], introduces a diffusion approximation for a Markov-modulated queue where the underlying Markov chain has two states. The approximation is based on matching the first and second moment of the workload process and of the approximating diffusion, but is not theoretically justified. The diffusion approximation leads to approximate *steady-state* performance metrics. On the other hand, the early works[7,23] were concerned with a theoretical analysis of Markov-modulated queueing systems. In particular[23], considers a single-server queue with Markov-modulated Poisson input and general service times which depend on the underlying Markov chain. They obtain expressions for the *steady-state* waiting time and queue length at arrival epochs and in continuous-time by matrix factorization methods applied to Wiener-Hopf-like equations[27]. considers single-server queue with K classes of customers and Markov-modulated arrivals. Their main contribution is allowing the service speed to depend on the underlying Markov chain. The main results of the article hinge on a time-change that normalizes the work speed to one. The authors then give relationships between performance measures of the original system and the normalized system. Crucially, most of their results concern the steady-state behavior of the queue. They characterize the distribution of the length of a busy period, and the number of customers served in a busy period. However, this is given in terms of a fixed-point equation which is not solved explicitly. Their only explicit solution is in steady-state. In Asmussen and Bladt[1], the authors derive an explicit expression for the mean busy period of a general Markov-modulated queue using a sample path approach. To do this, they generalize the classical relationship between the stationary distribution of a queue and the distribution of the maximum of the time-reversed net input process[21].

In Garikiparthi et al.[10], the authors consider a quasi-birth-death process and they derive the distribution of the number of customers served during a busy period. This process corresponds to a queueing system where the maximum system size is limited. Their model is quite different from ours. Indeed, the busy period distribution they find does not depend on how many busy periods have already occurred. Nevertheless, their techniques are loosely related to ours. They find a recursive expression for the matrix of conditional probabilities representing paths of height less than some level j and length $2i$. They write a recursive equation, over both i and j, for determining these matrices. However, they do not solve the recursion. They use a similar argument to justify a recursion for the joint Laplace transform of the number served and length of a busy period. In the works[9,11], the authors consider MEP/MEP/1 queueing systems, where MEP stands for Matrix Exponential Process, which includes the Poisson

process, and renewal processes. This allows both arrival and service processes to be non-renewal. The techniques are very similar to the ones in Garikiparthi et al.[10], i.e., they derive recursive equations for matrices summarizing quantities of interest (and their moments), even if these recursions are not solved explicitly. Both[9,11] remark on the combinatorial complexity of the sample path analysis, since also for their model there is a deep connection with Dyck paths. The likelihood of any such path, however, does not only depend on the number of upwards/downwards moves, but is crucially path-dependent – this is a key feature also of our model.

More recently[24], considers a single-server queue with Markov-modulated Poisson input and exponential service times. Their main contribution is allowing the service speed to depend on the queue length. Approximate expressions for steady-state performance measures are provided.

1.2. Organization

The rest of the article is organized as follows. In Section 2 we define the $\Delta_{(i)}/G/1$ model formally and we state our main result. In Section 3 we prove our main result by first developing a recursion for the distribution of the number of customers served in the first busy period, and then solving the recursion explicitly.

2. Model description, Dyck paths and main result

Consider a single-server queue that serves customers in a first-come first-served manner. There is a finite pool of N customers, each of which enters the system only once. Each customer independently joins the queue after an exponential time with rate λ and requires a service time that is exponentially distributed with rate μ. For notational convenience we denote by

$$\lambda_n := \lambda(N-n) \tag{1}$$

the arrival rate of customers to the system if n customers have already arrived to the system. In fact, our main result holds for *any* sequence of transition rates $(\lambda_n)_{n=1}^N$ such that $\lambda_i<\lambda_j$ for $i>j$, and such that $\lambda_N = 0$. However, when λ_n is not given by Eq. (1), the resulting Markov chain does not correspond to the $\Delta_{(i)}/G/1$ model. When λ_n is given by Eq. (1), the resulting combinatorial expressions simplify significantly, and so in our results throughout the article we will give both the general expressions, as well as the expressions corresponding to the specific choice Eq. (1). More generally, our technique also works for state-dependent service times $(\mu_n)_{n=1}^N$, as long as $\rho_j := \mu_j/(\mu_j+\lambda_j)$ is such that $\rho_1<\rho_2<\ldots<\rho_N = 1$. In this more general model, the service rate may depend on the number of

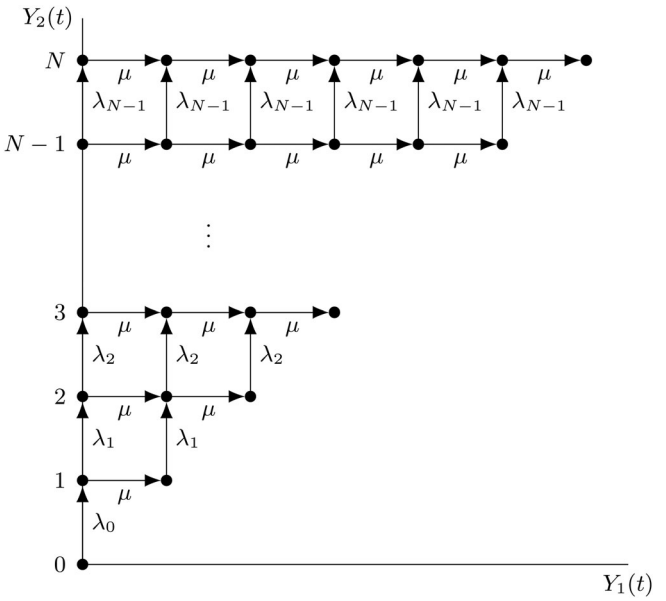

Figure 1. Transition rate diagram of the continuous-time Markov chain $\{Y(t)\}_{t\geq 0}$.

customers that have left the pool. For the sake of simplicity, in the rest of the article we only consider the case of constant service rate μ.

The state of the system at time $t \geq 0$ is described by a vector $Y(t) := (Y_1(t), Y_2(t)) \in \mathbb{N}^2$ where $Y_1(t)$ is the number of completed services at time t and $Y_2(t)$ is the number of customers that have joined the system up until time t. In view of our assumptions, the process $\{Y(t)\}_{t\geq 0}$ is a continuous-time Markov chain on the state space

$$\mathcal{S} := \{(i,j) \in \mathbb{N}^2 : 0 \leq j \leq N,\ \ 0 \leq i \leq j\}. \tag{2}$$

The transition rate diagram is depicted in Figure 1. The continuous-time Markov chain $\{Y(t)\}_{t\geq 0}$ is clearly reducible and admits the trivial equilibrium distribution π with $\pi_{N,N} = 1$ and $\pi_{i,j} = 0$ otherwise.

As illustrated in Figure 1, the state space $\mathcal{S}$ is highly structured. Our approach crucially leverages this structure. We refer to the set of states in the j-th row of $\mathcal{S}$

$$\mathcal{P}_j := \{(0,j), (1,j), ..., (j,j)\} \tag{3}$$

as the j-th *phase*, which corresponds to the situation in which exactly j customers have arrived in the system. We denote the collection of diagonal states as $\mathcal{D}_0 := \{(1,1), (2,2), ..., (N,N)\}$, and further use the notation $\mathcal{D}_n := \{(0,n), (1,n+1), ..., (N-n,N)\}$, $1 \leq n \leq N$ to denote the set of states on the n-th superdiagonal of $\mathcal{S}$.

It does not seem possible to find an explicit solution for the Kolmogorov equations associated to $Y(t)$ due to the time-inhomogeneous arrival process.

Therefore, we study the jump chain on $\mathcal{S}$ associated to $\{Y(t)\}_{t\geq 0}$, which we denote as $(X(k))_{k=0}^{2N}$. Conditionally on $X(k)=(i,j)$ with $i<j$, we have

$$X(k+1)=\begin{cases}(i+1,j) & \text{with probability } \rho_j\\ (i,j+1) & \text{with probability } 1-\rho_j,\end{cases} \tag{4}$$

where

$$\rho_j := \frac{\mu}{\mu+\lambda_j}. \tag{5}$$

In terms of the queueing system, ρ_j is the probability that a service occurs before an arrival when j customers have already arrived, but not all of them have already been served. Note that, conditionally on $X(k)=(i,i)$, we have $X(k+1)=(i,i+1)$ with probability one.

The $\Delta_{(i)}/G/1$ queueing model corresponds to the choice $\rho_j=0$ if $j=0$ and $\rho_j=\mu/(\mu+\lambda_j)$ if $j=1,\ldots,N$. We focus on the random variable S describing the number of customers served in the first busy period, which is the time between the instant a customer arrives to an empty system and the instant a customer departs the system leaving behind an empty system. Our main result is an explicit expression for the probability s_i that exactly i customers are served in the first busy period, i.e., $s_i := \mathbb{P}(S=i)$.

From the discussion above it follows that the trajectory of the Markov chain is a Dyck path of order N. For any $n\in\mathbb{N}_+$, we denote the set of Dyck paths of order n as $\mathfrak{D}_n$. A Dyck path $u\in\mathfrak{D}_n$ is fully characterized by the sequence $(u_j)_{j=1}^n$ of jumps to the right at each of the *phases* $\mathcal{P}_j$, with $j=1,\ldots,n$. With an abuse of notation we write

$$\mathfrak{D}_n=\{(u_1,\ldots,u_n)\in\mathbb{N}^n : \sum_{j=1}^{k} u_j \leq k \text{ for all } k=1,\ldots,n-1, \text{ and } \sum_{j=1}^{n} u_j = n\}. \tag{6}$$

The transition probabilities Eq. (4) of the Markov chain introduced above induce a probability measure $\bar{\mathbb{P}}$ on $\mathfrak{D}_N$ such that,

$$\bar{\mathbb{P}}(u)=\prod_{j=1}^{N}\rho_j^{u_j}(1-\rho_j)^{1\left\{\sum_{i=1}^{j}u_i<j\right\}}, \quad u=(u_1,\ldots,u_N)\in\mathfrak{D}_N. \tag{7}$$

From a probabilistic perspective, Eq. (7) can be understood as follows: the probability that the Markov chain jumps u_j times to the right at phase $\mathcal{P}_j$ is $\rho_j^{u_j}$. Moreover, if $\sum_{i=1}^{j}u_i=j$, then the Markov chain hits the diagonal on (j,j) and in that case it jumps up with probability one. Otherwise, it jumps up with probability $1-\rho_j$. From a combinatorial perspective, ρ_j and $1-\rho_j$ may be interpreted as *weights* associated to their respective edges

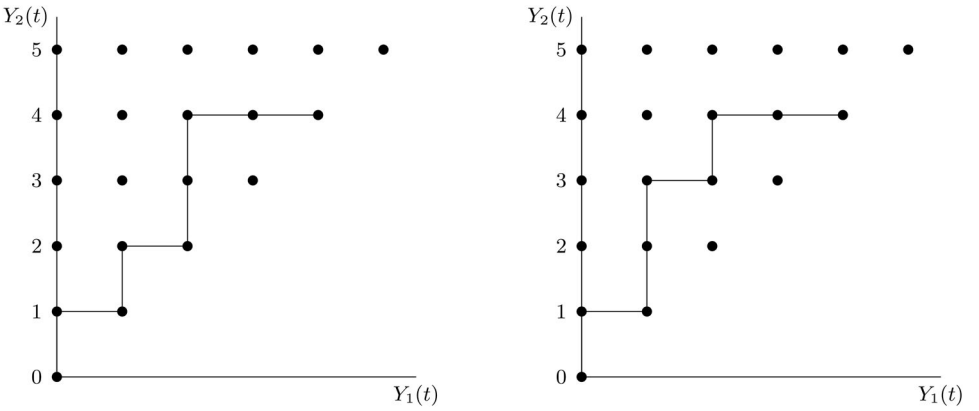

Figure 2. Examples of feasible and unfeasible allocations in $\mathcal{U}_4$ displayed as Dyck paths. The Dyck path on the left corresponds to the feasible allocation $u = (1, 1, 0, 2)$, the one on the right corresponds to the allocation $u = (1, 0, 1, 2)$, which is unfeasible since u_3 must be 0 or 2.

in $\mathcal{S}$. Equation (7) then assigns to the Dyck path u a weight $\bar{\mathbb{P}}(u)$, which is simply the product of the weights of the edges it traverses.

Equation (7) suggests partitioning the state space $\mathcal{S}$ in the N phases $\mathcal{P}_1, ..., \mathcal{P}_N$ in order to study the probability measure $\bar{\mathbb{P}}$. Crucially, the $(j+1)-$th phase may only be reached from the j-th phase and the transition probabilities between $\mathcal{P}_j$ and $\mathcal{P}_{j+1}$ only depend on j. We exploit this recursive structure by associating to each phase a generating function $P_j(z)$ and then expressing $P_{j+1}(z)$ in terms of $P_j(z)$. We then obtain the probability mass function of the number of customers served in the *first* busy period (equivalently, the probability mass function of the length of the first excursion of the associated Dyck path above the diagonal) by computing $P_j(\bar{z})$ for some explicit $\bar{z} \in \mathbb{R}$. We are able to fully solve this recursion by rewriting it as a linear system of equations and then inverting its coefficients matrix.

A crucial role in our result will be played by those Dyck paths that hit the diagonal whenever they jump to the right, see Figure 2. We make this precise in terms of the number of right jumps $(u_1, ..., u_n)$ of the Dyck path u at each phase $\mathcal{P}_j$. We define a *feasible allocation* $(u_1, ..., u_n)$ in a recursive manner, starting from u_1, as follows: u_1 is either 1 or 0, then

1. If $u_{i-1} = u_{i-2} = ... = u_{i-k+1} = 0$, and $u_{i-k} \neq 0$, then u_i is either k or 0;
2. If $u_{i-1} \neq 0$, then u_i is either 1 or 0.

Moreover, $(u_1, ..., u_n)$ is such that $\sum_{i=1}^{n} u_i = n$. We denote by $\mathcal{U}_n$ the set of feasible allocations. Equivalently,

$$\mathcal{U}_n := \{(u_1, ..., u_n) \in \mathfrak{D}_n : \forall 1 \leq i \leq n-1, \text{either } u_i = 0 \text{ or } \sum_{j=1}^{i} u_j = i\}.$$

With a minor abuse of terminology, we refer to elements of $\mathcal{U}_n$ interchangeabily as feasible allocations or as Dyck paths. The set $\mathcal{U}_n$ then

represents all those Dyck paths of order $n \leq N$ that hit the diagonal whenever they jump to the right. Some examples of feasible allocations for $n=4$ are $(1,1,0,2)$, $(0,0,3,1)$, $(0,2,0,2)$ and $(0,0,0,4)$. Some examples of unfeasible allocations for $n=4$ are $(1,0,1,2)$, since if $u_1=1$, $u_2=0$, then u_3 must be 0 or 2, $(1,0,0,2)$, since if $u_1=1$, $u_2=0$, and $u_3=0$, then u_4 must be 3, and $(0,0,2,2)$, since if $u_1=0$, $u_2=0$, then u_3 must be 0 or 3. See Figure 2 for an example of both a feasible and an unfeasible allocation in terms of Dyck paths. For every Dyck path $u \in \mathcal{U}_N$ there exists $\mathcal{J} = \mathcal{J}(u) \subseteq \{1,...,N\}$ such that Eq. (7) simply reads

$$\bar{\mathbb{P}}(u) = \prod_{j\in\mathcal{J}} \rho_j^{u_j} \prod_{k\in J^c} (1-\rho_k), \tag{8}$$

where $\mathcal{J}^c := \{1,...,N\} \setminus \mathcal{J}$. Here the set $\mathcal{J}$ represents the phases where the Dyck path jumps to the right and hits the diagonal. The set $\mathcal{J}^c$ then represents the phases where the Dyck path jumps up without jumping to the right. Conditioning on the phase in which the path first jumps to the right, it can be shown that $|\mathcal{U}_n| = 2^{n-1}$ for $1 \leq n \leq N$. In order to state our main result, we need an additional definition. For any $n \in \mathbb{N}$ and any vector $\boldsymbol{a} = (a_1,...,a_n) \in (\mathbb{R}^+)^n$, we define $M = M(\boldsymbol{a})$ to be the number of entries of the vector $\boldsymbol{a}$ that are not equal to one, i.e.,

$$M = M(\boldsymbol{a}) := \sum_{i=1}^{n} \mathbb{1}\{a_i \neq 1\}, \tag{9}$$

and by $j_{(1)}<j_{(2)}<...<j_{(M)}$ the ordered indices corresponding to those entries of $\boldsymbol{a}$. For notational convenience, we also define $j_{(0)} := 1 \leq j_{(1)}$ and $j_{(M+1)} := n \geq j_{(M)}$, so that

$$\boldsymbol{a} = (a_{j_{(0)}}, 1, ..., 1, a_{j_{(1)}}, 1, ..., 1, a_{j_{(M-1)}}, 1, ..., 1, a_{j_{(M)}}, 1, ..., 1, a_{j_{(M+1)}}).$$

We then introduce the function $b_n : (\mathbb{R}^+)^n \to \mathbb{R}$,

$$b_n(\boldsymbol{a}) := (-1)^{M-1} \prod_{m=0}^{M} \prod_{k=j_{(m)}}^{j_{(m+1)}-1} \frac{\lambda_k}{\lambda_k - \lambda_{j_{(m+1)}}}, \quad \boldsymbol{a} \in (\mathbb{R}^+)^n. \tag{10}$$

Note that plugging in $\lambda_n = \lambda(N-n)$ we get

$$b_n(\boldsymbol{a}) = (-1)^{M-1} \prod_{m=0}^{M} \binom{N-j_{(m)}}{j_{(m+1)}-j_{(m)}}. \tag{11}$$

Note that $b(\cdot)$ takes both positive and negative values. We can now state our main result.

Theorem 2.1. *The probability that exactly i customers are served in the first busy period of the $\Delta_{(i)}/G/1$ queue or, equivalently, the probability that the corresponding Dyck path hits the diagonal for the first time in (i, i) is*

given by

$$\mathbb{P}(S=i)=s_i=\sum_{(u_1,u_2,\ldots,u_i)\in\mathcal{U}_i} b_i(\rho_1^{u_1},\rho_2^{u_2},\ldots,\rho_i^{u_i})\rho_1^{u_1}\rho_2^{u_2}\cdots\rho_i^{u_i}, \tag{12}$$

where $b_i(\cdot)$ is given in Eq. *(10)*.

From a combinatorial perspective, s_i may be interpreted as the sum of the weights of all those Dyck paths of order i that do not hit the diagonal, which are in bijection with Dyck paths of order $i-1$. Then, equation (12) may be interpreted as a decomposition of the sum of weighted Dyck paths of order $i-1$ in terms of only those weighted Dyck paths that are associated with feasible allocations in $\mathcal{U}_i$ (the right-hand side). The term $b(\cdot)$ then represents the contribution of the upward jumps to the total weight of the path $u=(u_1,\ldots,u_i)$. The weight of each edge of the path depends on the phase where it is located, hence to compute the total weight of the path it is crucial to keep track of the location of the rightward jumps. This is accomplished by the indices $j_{(1)},\ldots,j_{(M)}$ associated to the Dyck path u. In particular, between the $j_{(m)}$-th phase and the $j_{(m+1)}$-th phase, u only makes upward jumps. With this in mind, M represents the total number of excursions above the diagonal that u makes.

Let us briefly make explicit the dependence of s_i on the initial number of customers N as $s_i^{(N)}$. Then, conditionally on $S=n$, the probability that i customers are served in the second busy period is $s_i^{(N-n)}$ and, hence, Theorem 2.1 gives the joint distribution of the number of customers served in *all* busy periods.

3. The number of customers in the first busy period

We prove Theorem 2.1 in two steps. First, in Subsection 3.1 we define a generating function $P_j(z)$ associated to phase j and derive a relation between $P_j(z)$ and $P_{j-1}(z)$. The probabilities s_i are obtained by evaluating $P_n(\bar{z})$ in a specific point $\bar{z}=\bar{z}(n)$, yielding a recursive relation for $s_1,\ldots,s_N$. Then, in Subsection 3.2 we interpret this recursive relation as a linear system $A\mathbf{s}=\mathbf{b}$, where $\mathbf{s}=(s_1,\ldots,s_N)$ and A is a lower-triangular matrix. By calculating the inverse A^{-1} explicitly, we finally obtain the expression for the probabilities $\mathbf{s}=(s_1,\ldots,s_N)$ as stated in Eq. (12).

3.1. Developing a recursion

We begin by introducing some notation. For every subset $\mathcal{A}\subsetneq\mathcal{S}$, the hitting time $H_{\mathcal{A}}$ is the random variable

$$H_{\mathcal{A}}:=\inf\{t\geq 0: Y(t)\in\mathcal{A}\}, \tag{13}$$

which describes the first time that the process $\{Y(t)\}_{t\geq 0}$ started at (0, 0)

enters the subset $\mathcal{A}$. When $x \in \mathcal{S}$ is a singleton, H_x should be understood as $H_{\{x\}}$.

Let $p_n(i)$ be the probability that the continuous-time Markov chain $\{Y(t)\}_{t\geq 0}$ first visits phase n hitting state (i, n) and without previously visiting $\mathcal{D}_0$, i.e.,

$$p_n(i) := \mathbb{P}\left(H_{\mathcal{P}_n} < H_{\mathcal{D}_0},\ X(H_{\mathcal{P}_n}) = (i,n)\right), \quad 0 \leq i \leq n,\ 1 \leq n \leq N. \tag{14}$$

Note that $p_n(n-1) = p_n(n) = 0$ for $2 \leq n \leq N$. Define the generating function of the sequence $(p_n(i))_{i=0}^{n-2}$ as

$$P_n(z) := \sum_{i=0}^{n-2} p_n(i) z^i, \quad z \in \mathbb{C},\ 2 \leq n \leq N. \tag{15}$$

For notational convenience, we also define $P_1(z) := 1$. Clearly, if $N = 1$, then $s_1 = 1$, hence from now on we will focus on $N > 1$. The strong Markov property implies that $s_1 = \rho_1$, and furthermore

$$s_n = \sum_{i=0}^{n-2} p_n(i) \rho_n^{n-i} = \rho_n^n P_n(\rho_n^{-1}), \quad 2 \leq n \leq N, \tag{16}$$

where ρ_n is defined in Eq. (5). Note that Eq. (16) implies $s_N = P_N(1)$. Equation (16) is the crucial relation that allows us to obtain a recursive expression for the probabilities $(s_n)_{n=1}^N$ starting from a recursive expression for the generating functions $(P_n(\cdot))_{n=1}^N$.

Finally, let $G_p(z)$ denote the probability generating function of a geometric random variable with support $\{0, 1, \ldots\}$ and success probability $1-p$, i.e.,

$$G_p(z) := \frac{1-p}{1-pz}, \quad |z| < \frac{1}{p}. \tag{17}$$

We are now ready to state our first result.

Lemma 3.1. *For any choice of positive transition probabilities $(\rho_j)_{j=1}^N$, the generating functions satisfy the recursion*

$$P_{n+1}(z) = G_{\rho_n}(z)[P_n(z) - s_n z^n], \quad 1 \leq n \leq N-1. \tag{18}$$

In particular,

$$P_{n+1}(z) = \prod_{i=1}^{n} G_{\rho_i}(z) - \sum_{i=1}^{n} s_i z^i \prod_{j=i}^{n} G_{\rho_j}(z), \quad |z| < \frac{1}{\rho_n}. \tag{19}$$

Proof. We start by expressing $P_{n+1}(z)$ in terms of $P_n(z)$. From the strong Markov property at time $H_{\mathcal{P}_n}$ we can write

$$p_{n+1}(i) = \sum_{j=0}^{i} p_n(j)\rho_n^{i-j}(1-\rho_n), \quad 0 \le i \le n-2, \tag{20}$$

$$p_{n+1}(n-1) = \sum_{j=0}^{n-2} p_n(j)\rho_n^{n-1-j}(1-\rho_n). \tag{21}$$

Multiply both sides of Eq. (20) by z^i and sum over all i with $0 \le i \le n-2$ and multiply both sides of Eq. (21) by z^{n-1}. Sum the two resulting expressions to get

$$P_{n+1}(z) = \sum_{i=0}^{n-2}\sum_{j=0}^{i} p_n(j)\rho_n^{i-j}(1-\rho_n)z^i + \sum_{j=0}^{n-2} p_n(j)\rho_n^{n-1-j}(1-\rho_n)z^{n-1}. \tag{22}$$

Switch the order of the double summation to obtain

$$\begin{aligned} P_{n+1}(z) &= (1-\rho_n)\left[\sum_{j=0}^{n-2} p_n(j)\sum_{i=j}^{n-2}\rho_n^{i-j}z^i + \sum_{j=0}^{n-2} p_n(j)\rho_n^{n-1-j}z^{n-1}\right] \\ &= (1-\rho_n)\left[\sum_{j=0}^{n-2} p_n(j)\sum_{k=0}^{n-2-j}\rho_n^{k}z^{j+k} + \sum_{j=0}^{n-2} p_n(j)\rho_n^{n-1-j}z^{n-1}\right]. \end{aligned} \tag{23}$$

The summation over k is a geometric sum. Performing this summation and rewriting yields the recursive expression

$$\begin{aligned} P_{n+1}(z) &= (1-\rho_n)\left[\sum_{j=0}^{n-2} p_n(j)\frac{z^j-\rho_n^{n-1-j}z^{n-1}}{1-\rho_n z} + \sum_{j=0}^{n-2} p_n(j)\rho_n^{n-1-j}z^{n-1}\right] \\ &= \frac{1-\rho_n}{1-\rho_n z}\left[\sum_{j=0}^{n-2} p_n(j)z^j - z^n\rho_n^n\sum_{j=0}^{n-2} p_n(j)\rho_n^{-j}\right] \\ &= G_{\rho_n}(z)[P_n(z)-s_n z^n]. \end{aligned} \tag{24}$$

To prove the explicit expression Eq. (19) we iterate the recursion Eq. (24), obtaining

$$P_{n+1}(z) = P_2(z)\prod_{i=2}^{n} G_{\rho_i}(z) - \sum_{i=2}^{n} s_i z^i \prod_{j=i}^{n} G_{\rho_j}(z), \tag{25}$$

which we can further simplify by noting that

$$P_2(z) = p_2(0) = 1-\rho_1 = (1-\rho_1 z)G_{\rho_1}(z). \tag{26}$$

Since $s_1 = \rho_1$, we finally obtain Eq. (19).

Note that for the $\Delta_{(i)}/G/1$ queue we have $\rho_i^{-1} = (\mu + \lambda_i)/\mu > (\mu + \lambda_j)/\mu = \rho_j^{-1}$ for any $i < j$. Therefore, $G_{\rho_i}(\rho_n^{-1})$ is well defined for all $i < n$.

In the proof of Lemma 3.1 we did not make use of the precise expression of ρ_n, and so Eq. (19) still holds when replacing λ_i with any sequence of positive decreasing numbers. Combining Lemma 3.1 with Eq. (16) allows us to obtain a recursive expression for s_n. We first present the expression for s_n for a general decreasing sequence $(\lambda_n)_{n=1}^N$, and then the one obtained when setting $\lambda_n = \lambda(N-n)$. We adopt the convention that the empty sum $\sum_{i=1}^0(\cdot) = 0$ and the empty product $\prod_{i=1}^0(\cdot) = 1$.

Corollary 3.2. *Assume $(\lambda_n)_{n=1}^N$ is a sequence such that $\lambda_1 > \ldots > \lambda_{N-1} > \lambda_N = 0$. Then,*

$$s_n = \rho_n^n \prod_{k=1}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n} - \sum_{i=1}^{n-1} s_i \rho_n^{n-i} \prod_{k=i}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n}, \quad 2 \le n \le N, \tag{27}$$

with initial term $s_1 = \rho_1$. In particular, when $\lambda_n = \lambda(N-n)$, the probabilities s_n satisfy the recursion

$$s_n = \rho_n^n \binom{N-1}{n-1} - \sum_{i=1}^{n-1} s_i \rho_n^{n-i} \binom{N-i}{n-i}, \tag{28}$$

with initial term $s_1 = \rho_1$.

Proof. Combining the result of Lemma 3.1 with Eq. (16) yields the following recursion, for $2 \le n \le N-1$,

$$s_n = \rho_n^n \prod_{i=1}^{n-1} G_{\rho_i}(\rho_n^{-1}) - \sum_{i=1}^{n-1} s_i \rho_n^{n-i} \prod_{j=i}^{n-1} G_{\rho_j}(\rho_n^{-1}), \quad s_N = 1 - \sum_{i=1}^{N-1} s_i. \tag{29}$$

Note that, by our assumption on the sequence $(\lambda_n)_{n=1}^N$, we have $\rho_1^{-1} > \cdots > \rho_N^{-1} = 1$. Therefore, $G_{\rho_i}(\rho_n^{-1})$ is well defined for all $i < n$. The first expression Eq. (27) follows from

$$G_{\rho_k}(\rho_n^{-1}) = \frac{1-\rho_k}{1-\frac{\rho_k}{\rho_n}} = \frac{1-\frac{\mu}{\mu+\lambda_k}}{1-\frac{\mu+\lambda_n}{\mu+\lambda_k}} = \frac{\lambda_k}{\lambda_k - \lambda_n}. \tag{30}$$

Moreover, when $\lambda_n = \lambda(N-n)$ we get

$$\prod_{k=l}^{n-1} G_{\rho_k}(\rho_n^{-1}) = \prod_{k=l}^{n-1} \frac{N-k}{n-k} = \frac{N-l}{n-l} \frac{N-l-1}{n-l-1} \frac{N-l-2}{n-l-2} \cdots \frac{N-n+1}{1} = \binom{N-l}{n-l},$$

which proves Eq. (28).

3.2. Solving the recursion

In this section we solve the recursion Eq. (27) to find an explicit expression for s_n. Recall that for $n = 1, 2, \ldots, N$,

$$s_n = \rho_n^n \prod_{k=1}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n} - \sum_{i=1}^{n-1} s_i \rho_n^{n-i} \prod_{k=i}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n}, \tag{31}$$

Divide both sides by ρ_n^n and bring all s_i terms to one side to obtain

$$\sum_{i=1}^{n} \frac{s_i}{\rho_n^i} \prod_{k=i}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n} = \prod_{k=1}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n}. \tag{32}$$

We can write Eq. (32) in the matrix-vector notation $A\mathbf{s} = \mathbf{b}$, where we introduced the column vectors

$$\mathbf{s} := (s_i)_{i=1,2,\ldots,N}, \text{ and } \mathbf{b} := \left(\prod_{k=1}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n} \right)_{n=1,2,\ldots,N} \tag{33}$$

and the lower-triangular matrix A with element (n, i) given by

$$(A)_{n,i} := \frac{1}{\rho_n^i} \prod_{k=i}^{n-1} \frac{\lambda_k}{\lambda_k - \lambda_n}, \quad 1 \le i \le n \le N. \tag{34}$$

We can calculate $\mathbf{s}$ as $\mathbf{s} = A^{-1}\mathbf{b}$. In particular, since A is a lower-triangular matrix, so is its inverse A^{-1}. Hence, we can determine the inverse using the well-known recursive formulas

$$(A^{-1})_{n,n} = \frac{1}{(A)_{n,n}} = \rho_n^n, \quad n = 1, 2, \ldots, N, \tag{35}$$

$$(A^{-1})_{n,i} = -(A^{-1})_{i,i} \sum_{k=i+1}^{n} (A^{-1})_{n,k} (A)_{k,i}, \quad 1 \le i < n \le N. \tag{36}$$

This recursion is solved in a specific order. One first determines $(A^{-1})_{n,n}$, for $n = 1, 2, \ldots, N$, then all $(A^{-1})_{n,n-1}$, for $n = 2, 3, \ldots, N$, followed by $(A^{-1})_{n,n-2}$, for $n = 3, 4, \ldots, N$, and so on until finally $(A^{-1})_{N,1}$ is reached. We exploit this recursion in order to derive an explicit expression for the elements of the inverse.

To this end, it is useful to work with a slightly more general definition of the function $b(\cdot)$ introduced in Eq. (10). Let $\mathcal{S}_n := \{(k_1, \ldots, k_n) \in \mathbb{N}^n : k_1 < k_2 < \ldots < k_n\}$ be the set of non-decreasing indices of length n. Recall that $M = M(\boldsymbol{a})$ is the number of entries of the vector $\boldsymbol{a}$ that are not equal to one and $j_{(1)} < j_{(2)} < \ldots < j_{(M)}$ are ordered indices corresponding to those entries. For a vector $\boldsymbol{a} \in (\mathbb{R}^+)^n$ and a vector of indices $\boldsymbol{k} \in \mathcal{S}_n$, we define $k_{(0)} := k_1$, $k_{(1)} := k_{j_{(1)}}$, …, $k_{(M)} := k_{j_{(M)}}$, and $k_{(M+1)} :=$

k_n. We introduce the function $\tilde{b}_n : (\mathbb{R}^+)^n \times \mathcal{S}_n \mapsto \mathbb{R}$ that associates to the vector $\boldsymbol{a} = (a_1, ..., a_n)$ and a set of indices $\boldsymbol{k} = (k_1, ..., k_n)$ the scalar $\tilde{b}_n(\boldsymbol{a}, \boldsymbol{k})$ defined as

$$\tilde{b}_n(\boldsymbol{a}, \boldsymbol{k}) := (-1)^{M-1} \prod_{m=0}^{M} \prod_{k=k_{(m)}}^{k_{(m+1)}-1} \frac{\lambda_k}{\lambda_k - \lambda_{k_{(m+1)}}}. \tag{37}$$

Note that $b_n(\cdot)$ in Eq. (10) is recovered when $\boldsymbol{k} = (1, 2, ..., n)$, since

$$b_n(\boldsymbol{a}) = \tilde{b}_n(\boldsymbol{a}, (1, 2, ..., n)). \tag{38}$$

In order to prove Theorem 2.1, we first obtain an explicit expression for the inverse coefficient matrix A^{-1}.

Proposition 3.3. *Assume that $(\lambda_n)_{n=1}^N$ is a sequence such that $\lambda_1 > ... > \lambda_{N-1} > \lambda_N = 0$. Then, for any $i = 1, ..., N$ and $n = 1, 2, ..., i-1$ we have*

$$(A^{-1})_{i, i-n} = \sum_{(u_1, u_2, ..., u_n) \in \mathcal{U}_n} \tilde{b}_n(\boldsymbol{\rho}^u_{i,n}, \boldsymbol{k}_{i,n}) \rho_{i-n}^{i-n} \rho_{i-n+1}^{u_1} \rho_{i-n+2}^{u_2} \cdots \rho_i^{u_n}, \tag{39}$$

where

$$\boldsymbol{\rho}^u_{i,n} := (\rho_{i-n}^{i-n}, \rho_{i-n+1}^{u_1}, ..., \rho_i^{u_n}), \tag{40}$$

$$\boldsymbol{k}_{i,n} := (i-n, i-n+1, ..., i), \tag{41}$$

and $\tilde{b}(\cdot)$ was defined in Eq. (37).

Proof. We proceed by induction, by assuming that Eq. (39) holds for all $m \leq n$ for some $n \in \{1, ..., i-1\}$ and then proving it for $n+1$. We use Eq. (36) together with Eq. (34) to obtain

$$\begin{aligned}
(A^{-1})_{i, i-(n+1)} &= -\rho_{i-(n+1)}^{i-(n+1)} \sum_{k=i-n}^{i} (A^{-1})_{i,k} (A)_{k, i-n-1} \\
&= -\rho_{i-(n+1)}^{i-(n+1)} \sum_{j=0}^{n} (A^{-1})_{i, i-j} (A)_{i-j, i-n-1} \\
&= -\rho_{i-(n+1)}^{i-(n+1)} \sum_{j=0}^{n} (A^{-1})_{i, i-j} \frac{1}{\rho_{i-j}^{i-(n+1)}} \prod_{k=i-(n+1)}^{i-j-1} \frac{\lambda_k}{\lambda_k - \lambda_{i-j}}.
\end{aligned} \tag{42}$$

In the last equality we highlight the inductive structure in the product term. To avoid encumbering the computations, let us denote the product in Eq. (42) as

$$\mathcal{B}_{i,j,n} := - \prod_{k=i-(n+1)}^{i-j-1} \frac{\lambda_k}{\lambda_k - \lambda_{i-j}}. \tag{43}$$

Inserting the expression for $(A^{-1})_{i,i-j}$ into Eq. (42) gives

$$\begin{aligned}
&\rho_{i-(n+1)}^{i-(n+1)}\sum_{j=0}^{n}(A^{-1})_{i,i-j}\frac{1}{\rho_{i-j}^{i-n-1}}\mathcal{B}_{i,j,n}\\
&\quad=\rho_{i-(n+1)}^{i-(n+1)}\sum_{j=0}^{n}\sum_{(u_1,\ldots,u_j)\in\mathcal{U}_j}\rho_{i-j}^{i-j}\rho_{i-j+1}^{u_1}\ldots\rho_{i}^{u_j}\tilde{b}_j(\boldsymbol{\rho}_{i,j}^{u},\boldsymbol{k}_{i,j})\frac{1}{\rho_{i-j}^{i-n-1}}\mathcal{B}_{i,j,n} \qquad (44)\\
&\quad=\sum_{j=0}^{n}\sum_{(u_1,\ldots,u_j)\in\mathcal{U}_j}\rho_{i-(n+1)}^{i-(n+1)}\rho_{i-j}^{n+1-j}\rho_{i-j+1}^{u_1}\ldots\rho_{i}^{u_j}\tilde{b}_j(\boldsymbol{\rho}_{i,j}^{u},\boldsymbol{k}_{i,j})\mathcal{B}_{i,j,n},
\end{aligned}$$

where, recall,

$$\boldsymbol{\rho}_{i,j}^{u}=(\rho_{i-j}^{i-j},\rho_{i-j+1}^{u_1},\ldots,\rho_{i}^{u_j}). \qquad (45)$$

Now, observe that $(n+1-j)+u_1+\ldots+u_j=n+1$. Crucially, we also have that

$$\begin{aligned}
&\sum_{j=0}^{n}\sum_{(u_1,\ldots,u_j)\in\mathcal{U}_j}\rho_{i-(n+1)}^{i-(n+1)}\rho_{i-j}^{n+1-j}\rho_{i-j+1}^{u_1}\ldots\rho_{i}^{u_j}\tilde{b}_j(\boldsymbol{\rho}_{i,j}^{u},\boldsymbol{k}_{i,j})\mathcal{B}_{i,j,n}\\
&\quad=\sum_{(v_1,\ldots,v_{n+1})\in\mathcal{U}_{n+1}}\rho_{i-(n+1)}^{i-(n+1)}\rho_{i-n}^{v_1}\ldots\rho_{i}^{v_{n+1}}\tilde{b}_{n+1}(\boldsymbol{\rho}_{i,(n+1)}^{v},\boldsymbol{k}_{i,n+1}),
\end{aligned} \qquad (46)$$

with

$$\begin{aligned}
\boldsymbol{\rho}_{i,(n+1)}^{v}&:=(\rho_{i-(n+1)}^{i-(n+1)},\rho_{i-n}^{v_1},\ldots,\rho_{i}^{v_{n+1}})\\
\boldsymbol{k}_{i,n+1}&:=(i-(n+1),i-n,\ldots,i).
\end{aligned}$$

Indeed, the left-hand side of Eq. (46) corresponds to the feasible assignment in which the first jump to the right occurs at phase $i-(n+1)$, which is necessarily of length $i-(n+1)$. Then, for any fixed $j=0,\ldots,n$, the next jump to the right occurs at phase $i-j$, which is necessarily of length $n+1-j$. A sum is then performed over the remaining feasible assignments. Summing over all possible $j=0,\ldots,n$ on the left-hand side of Eq. (46), one obtains a sum over all feasible assignments such that the first jump to the right occurs at phase $i-(n+1)$, which is the sum on the right-hand side of Eq. (46). Furthermore, for the vector $\boldsymbol{\rho}_{i,n+1,j}:=(\rho_{i-(n+1)}^{i-(n+1)},\rho_{i-j}^{n+1-j},\rho_{i-j+1}^{u_1},\ldots,\rho_{i}^{u_j})$ and the indices $\boldsymbol{k}_{i,n+1,j}:=(i-(n+1),i-j,i-j+1,\ldots,i)$, we have that $k_{(0)}=k_{(1)}=i-(n+1)$ and $k_{(2)}=i-j$, so that

$$\mathcal{B}_{i,j,n}=-\prod_{k=i-(n+1)}^{(i-j)-1}\frac{\lambda_k}{\lambda_k-\lambda_{i-j}}=-\prod_{k=k_{(1)}}^{k_{(2)}-1}\frac{\lambda_k}{\lambda_k-\lambda_{k_{(2)}}}. \qquad (47)$$

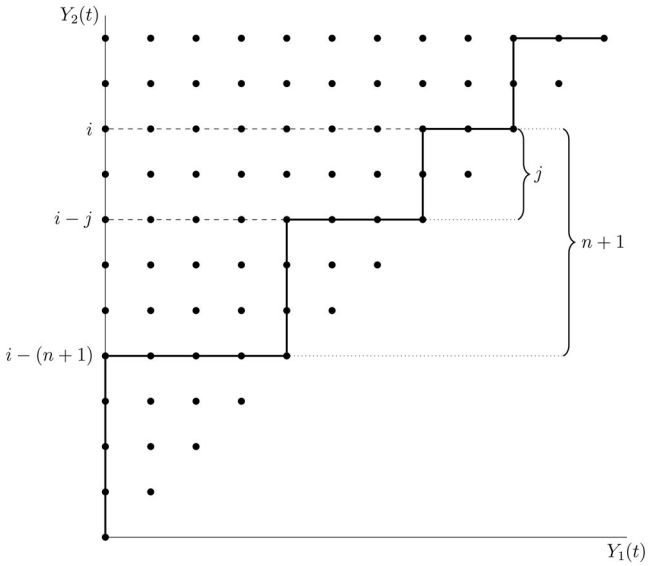

Figure 3. On the left-hand side of the inductive step Eq. (46), the first right jump of the Dyck path occurs at phase $i-(n+1)$, and the first jump after that occurs at phase $i-j$. Summing over $j=0,\ldots,n$, one obtains all paths that jump to the right for the first time at phase $i-(n+1)$, which is the right-hand side of Eq. (46).

It follows that

$$\tilde{b}_j(\boldsymbol{\rho}^u_{i,j},\boldsymbol{k}_{i,j})\mathcal{B}_{i,j,n}=\tilde{b}_{j+1}(\boldsymbol{\rho}_{i,n+1,j},\boldsymbol{k}_{i,n+1,j}). \tag{48}$$

Figure 3 illustrates this decomposition in terms of Dyck paths.

We can finally prove Theorem 2.1 by applying Proposition 3.3 to invert the matrix A.

Proof of Theorem 2.1. Writing $\mathbf{s}=A^{-1}\mathbf{b}$ explicitly yields

$$s_i=\sum_{n=0}^{i-1}(A^{-1})_{i,i-n}\prod_{k=1}^{i-n-1}\frac{\lambda_k}{\lambda_k-\lambda_{i-n}}. \tag{49}$$

Plugging Eq. (39) into Eq. (49), using the same inductive argument as in Eq. (46) and noting that

$$\prod_{k=1}^{i-n-1}\frac{\lambda_k}{\lambda_k-\lambda_{i-n}}=\prod_{k=k_{(0)}}^{k_{(1)}-1}\frac{\lambda_k}{\lambda_k-\lambda_{i-n}}, \tag{50}$$

gives

$$s_i=\sum_{(u_1,u_2,\ldots,u_i)\in\mathcal{U}_i}\rho_1^{u_1}\rho_2^{u_2}\cdots\rho_i^{u_i}\tilde{b}_i((\rho_1^{u_1},\rho_2^{u_2},\ldots,\rho_i^{u_i}),(1,2\ldots,i)), \tag{51}$$

concluding the proof.

ORCID

Gianmarco Bet http://orcid.org/0000-0001-8431-0636

References

[1] Asmussen, S.; Bladt, M. A sample path approach to mean busy periods for Markov-modulated queues and fluids. *Adv. Appl. Probab.* 1994, *26*, 1117–1121. DOI: 10.2307/1427907.

[2] Bet, G. An alternative approach to heavy-traffic limits for finite-pool queues. *Queue. Syst.* 2020, *95*, 121–144. DOI: 10.1007/s11134-020-09653-z.

[3] Bet, G.; van der Hofstad, R.; van Leeuwaarden, J. S. H. Finite-pool queueing with heavy-tailed services. *J. Appl. Probab.* 2017, *54*, 921–942. DOI: 10.1017/jpr.2017.42.

[4] Bet, G.; van der Hofstad, R.; van Leeuwaarden, J. S. H. Heavy-traffic analysis through uniform acceleration of queues with diminishing populations. *Math. Operat. Res.* 2019, *44*, 821–864. DOI: 10.1287/moor.2018.0947.

[5] Bet, G.; van der Hofstad, R.; van Leeuwaarden, J. S. H. Big jobs arrive early: From critical queues to random graphs. *Stochastic Syst.* 2020, *10*, 310–334. DOI: 10.1287/stsy.2019.0057.

[6] Brown, L.; Gans, N.; Mandelbaum, A.; Sakov, A.; Shen, H.; Zeltyn, S.; Zhao, L. Statistical analysis of a telephone call center. *J. Am. Stat. Assoc.* 2005, *100*, 36–50. DOI: 10.1198/016214504000001808.

[7] Burman, D. Y.; Smith, D. R. Asymptotic analysis of a queueing model with bursty traffic. *Bell Syst. Tech. J.* 1983, *62*, 1433–1453. DOI: 10.1002/j.1538-7305.1983.tb03490.x.

[8] Courtiel, J.; Melczer, S.; Mishna, M.; Raschel, K. Weighted lattice walks and universality classes. *J. Combinator. Theory, Ser. A.* 2017, *152*, 255–302. DOI: 10.1016/j.jcta.2017.06.008.

[9] Garikiparthi, C. *Sample Path Analysis of Stochastic Processes: Busy Periods of Auto-Correlated Single Server Queues*; University of Missouri-Kansas City: Kansas City, 2008.

[10] Garikiparthi, C.; van de Liefvoort, A.; Mitchell, K. Busy period analysis of finite QBD processes. *SIGMETRICS Perform. Eval. Rev.* 2008, *36*, 98–100. DOI: 10.1145/1453175.1453196.

[11] Garikiparthivan de Liefvoort, C. A.; Mitchell, K. Sample path analysis of busy periods and related first passages of a correlated MEP/MEP/1 system. In *Fourth International Conference on the Quantitative Evaluation of Systems (QEST 2007),* IEEE, 2007; 277–286.

[12] M.; Hanly, T.; Churches, O.; Fitzgerald, I.; Caterson, Chandini Raina MacIntyre, L. Jorm, Modelling vaccination capacity at mass vaccination hubs and general practice clinics; *medRxiv*, 2021. https://www.medrxiv.org/content/10.1101/2021.04.07.21255067v1.

[13] Haug, N.; Prellberg, T. Uniform asymptotics of area-weighted Dyck paths. *J. Math. Phys.* 2015, *56*, 043301. DOI: 10.1063/1.4917052.

[14] Honnappa, H.; Jain, R. Strategic arrivals into queueing networks: The network concert queueing game. *Oper. Res.* 2015, *63*, 247–259. DOI: 10.1287/opre.2014.1338.

[15] Honnappa, H.; Jain, R.; Ward, A. R. A queueing model with independent arrivals, and its fluid and diffusion limits. *Queue. Syst.* 2015, *80*, 71–103. DOI: 10.1007/s11134-014-9428-4.

[16] Honnappa, H.; Ward, A. R. *On transitory queueing*; 2014. https://arxiv.org/abs/1412.2321.

[17] Kaandorp, G. C.; Koole, G. Optimal outpatient appointment scheduling. *Health Care Manage. Sci.* 2007, *10*, 217–229. DOI: 10.1007/s10729-007-9015-x.

[18] Kim, S.-H.; Whitt, W. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manuf. Serv. Oper. Manage.* 2014, *16*, 464–480. DOI: 10.1287/msom.2014.0490.

[19] Kim, S.-H.; Whitt, W. Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Res. Logist.* 2014, *61*, 66–90. DOI: 10.1002/nav.21568.

[20] Louchard, G. Random walks, Gaussian processes and list structures. *Theoret. Comp. Sci.* 1987, *53*, 99–124. DOI: 10.1016/0304-3975(87)90028-4.

[21] Loynes, R. M. The stability of a queue with non-independent inter-arrival and service times. *Math. Proc. Camb. Phil. Soc.* 1962, *58*, 497–520. DOI: 10.1017/S0305004100036781.

[22] Parlar, M.; Sharafali, M. Dynamic allocation of airline check-in counters: A queueing optimization approach. *Manage. Sci.* 2008, *54*, 1410–1424. DOI: 10.1287/mnsc.1070.0842.

[23] Regterschot, G. J. K.; De Smit, J. H. A. The queue M/G/1 with Markov modulated arrivals and services. *Math. Oper. Res.* 1986, *11*, 465–483. DOI: 10.1287/moor.11.3.465.

[24] Sakthi, R.; Vidhya, V.; Mahaboob Sherieff, K. H. Performance measures of state dependent MMPP/M/1 queue. *IJET.* 2018, *7*, 942–945. DOI: 10.14419/ijet.v7i4.10.26632.

[25] Takahashi, H.; Akimaru, H. A diffusion model for queues in randomly varying environment. *IEICE Trans.* 1976–1990, *69*, 1, 13–20.

[26] Takahashi, H.; Wang, L.-S. An approximate analysis of a queueing system with Markov-modulated arrivals. *Electron. Comm. Jpn. (Part I: Commun.)* 1990, *73*, 12–21. DOI: 10.1002/ecja.4410731102.

[27] Takine, T. Single-server queues with Markov-modulated arrivals and service speed. *Queue. Syst.* 2005, *49*, 7–22. DOI: 10.1007/s11134-004-5553-9.

[28] Van Rensburg, E. J. J.; Prellberg, T.; Rechnitzer, A. Partially directed paths in a wedge. *J. Combinator. Theory, Ser. A.* 2008, *115*, 623–650. DOI: 10.1016/j.jcta.2007.08.003.

[29] Viennot, G. *A Combinatorial Theory for General Orthogonal Polynomials with Extensions and Applications, Polynômes Orthogonaux et Applications*; Springer: New York, 1985; 139–157.

[30] Wang, R.; Jouini, O.; Benjaafar, S. Service systems with finite and heterogeneous customer arrivals. *Manuf. Serv. Oper. Manage.* 2014, *16*, 365–380. DOI: 10.1287/msom.2014.0481.

[31] Whitt, W. *Stochastic-Process Limits. An Introduction to Stochastic-Process Limits and Their Application to Queues*; Springer: New York, 2002.

[32] Wang, R.; Jouini, O.; Benjaafar, S. *Queues with time-varying arrival rates: A bibliography*; http://www.columbia.edu/ww2040/TV_bibliography_091016.pdf. 2016.