On queues and numbers

Eitan Bachmat

July 1, 2012

1 Introduction

Queueing theory deals, as the name suggests, with queues. Typical examples can be queues that we encounter in the supermarket, the bank, a retail store or an amusement park. Other queues which are less directly visible but which certainly affect our experiences in the virtual world are queues in processors and over communication channels. The analysis of queues in queueing theory is based on making stochastic assumptions about the behavior of customers and service providers in the system. The resulting analysis leads to estimates or calculations of quantities such as the average waiting time of a class of customers in the system, or the probability that a customer will wait more than a certain amount of time.

Number theory seems to be very far removed from the world of queueing theory. The purpose of this article is to explain how a certain set of symmetries plays a role in both queueing theory and in number theory. The basic explanation for the role the symmetries play in both cases comes from a simple observation of Riemann.

We have chosen to introduce the subject by looking at a very specific problem. Analyzing the problem and putting it in context will lead us to consider some basic results in queueing theory but will also lead us in parallel to consider some basic results in number theory. While the problem we will study has been considered mostly in the context of computer servers, it will be convenient to explain it in the context of managing a mini-market with two checkout counters, one of them acting as an express line.

2 Managing a mini-market

2.1 The ingredients of the mini-market queueing system

We consider a mini-market with two checkout counters, numbered 1 and 2. In order to mathematically analyze a queueing system we need some information that we now specify. We first consider the process in which customers arrive to the queueing system.

We say that a queueing system has **Poisson arrivals** with rate λ if customers may arrive after some time $T = 0$ and the number of customers arriving at any given time interval $I = [a, b]$ has a Poisson distribution with parameter $\lambda(b - a)$, with arrivals in disjoint time intervals being independent.

Another way of describing Poisson arrivals is to say that the times between the successive arrivals of customers to the queue are independent of each other and exponentially distributed, i.e., $Pr(X_i \geq t) = e^{-\lambda t}$, where X_i is the time between the arrival of the customer i and customer $i + 1$.

Our mini-market has two checkout counters where service is provided. Today some mini-markets have automatic checkout counters, but we will deal with the old fashioned kind where there are actual people.

We will call the two persons working at the counters Alice and Bob. We will think of the process of handling a single customer as a **job**. We also assume that at each checkout counter, customers are serviced in the order in which they arrive. This is known as a `FIFO`, First In First Out, queue.

We will assume that each job (customer) which takes Alice t time to process, takes Bob ct time, where c is a constant which is independent of the job. Without loss of generality we can assume that Alice is the faster worker, i.e., $c \geq 1$. The case $c = 1$ corresponds to equally capable service providers and is also known as the **identical server** case.

Assume that in our queueing system, the job-size of the i'th customer with respect to Alice is a random variable X_i which is independent of any other random variable which defines the queueing process and that all X_i have a common distribution function $X(t) = Pr(X_i \geq t)$. Under these assumptions, the function X is known as the **job-size distribution**.

To define an express line we will need a processing time cutoff s. The cutoff s will always be with respect to the processing time of Alice. We will send all customers whose checkout processing time T satisfies $T \leq s$ to checkout counter 1 while those with processing time $T > s$ will be sent to checkout counter 2. It may be argued that we usually do not know the processing time T of a customer, but to simplify matters we will assume that we do know.

Since Alice and Bob are not equally capable in general, we have to decide who serves customers in the express checkout counter which is counter number 1. We will describe this decision by a permutation σ on two elements. The arrangement whereby Alice serves customers in the express counter and Bob is assigned to the other counter will be assigned to the identity permutation, while the other assignment in which Alice and Bob switch counters will correspond to the non-identity permutation. If σ is a permutation, then we denote by $\bar{\sigma}$ the reverse order permutation.

The parameters λ,X,s,c,σ, specify completely a mini-market queueing system, however, it is more convenient to replace the customer arrival rate λ by another parameter which takes into account the processing power of the servers.

The **utilization** ρ of a queueing system with a single server, arrival rate λ and job size distribution X is defined as $\rho = \lambda E(X)$, where $E(X)$ denotes the expectation of X. For the mini-market system described above we define the utilization as $\rho = \frac{\lambda E(X)}{1+1/c}$.

In a single server system the utilization tells us the portion of time in which the system is busy serving customers. We recall that we are assuming an arrival rate of λ customers per time unit on average. Since Alice has speed 1, the average size of a customer's processing time, w.r.t. Alice, is $E(X)$, the average of the job-size distribution X. Similarly, on average a job takes Bob $cE(X)$ time to process. This means that Alice can process on average $1/E(X)$ customers per time unit, while Bob can handle on average $1/cE(X)$ customers. Since they are working in parallel, they can handle up to $(1 + 1/c)/E(X)$ customers on average per time unit. If the customer arrival rate λ is larger than the average service rate $(1+1/c)/E(X)$, then regardless of the cutoff value, the queue at one of the counters will explode in size leading to an unstable system with infinitely large average waiting time. If, on the other hand, the ratio of customer arrival rate to customer service rate $\rho = \frac{\lambda E(X)}{1+1/c}$ satisfies $\rho < 1$, then, we can find a cutoff s for which the mini-market queueing system will stabilize as a stochastic process. In particular the average waiting time of a customer in the system after operating for T time units will tend with probability approaching 1 to a well defined limit as $T \to \infty$, see [26] or any other reference book on queueing theory for more details.

2.2 The mini-market management problem

We will state three management problems for a mini-market queueing system with an express counter as described above.

Problem 1: Assume that Alice and Bob are equally capable workers, i.e., $c = 1$. In this case the permutation σ is immaterial. Let $E(W)(X, \rho, s)$ be the average waiting time of customers in the mini-market queueing system (customer service time at the counter not included).

Given a pair X, ρ, define $E(W)_{opt} = Min_s \; E(W)(X, \rho, s)$. It is known that the minimum actually exists. We let $s_{opt} = s_{opt}(X, \rho)$ be an optimal cutoff value, namely, $E(W)(X, \rho, s_{opt}) = E(W)_{opt}$.

The problem is to compute the value of the optimal cutoff $s_{opt}(X, \rho)$.

Problem 2: We assume that $c \neq 1$, i.e., Alice and Bob are not equally capable. Fixing X, ρ, c, let $E(W)_{opt}(\sigma) = Min_s \; E(W)(X, \rho, c, \sigma, s)$.

The question is whether $E(W)_{opt}(\sigma) = E(W)_{opt}(\bar{\sigma})$. Stated otherwise, given X, ρ, c, does it matter whether the faster worker (Alice) works at the express counter or the slower worker (Bob)?

In case it matters we can also ask which assignment should be preferred, should we place the faster worker, Alice, at the express (small jobs) counter or the other way around?

Problem 3: Continuing the scenario of problem 2, for fixed parameters X, ρ, c, define $s_{opt}(\sigma)$ to be a cutoff for which $E(W)_{opt}(X, \rho, c, \sigma, s) = E(W)(X, \rho, c, \sigma, s_{opt})$. Suppose we know $s_{opt}(\sigma)$, can we easily compute $s_{opt}(\bar{\sigma})$ from it? In other words, suppose Alice has been working at the express counter for a while and that during this time the mini-market manager has been able to determine s_{opt} via trial and error or some other method. Suppose that one day Alice and Bob get bored and decide to switch counters, can the manager compute the new optimal cutoff from the old one?

After all this build up, we are sorry to report that for general parameters X, ρ, c, the answers to these problems are tedious or take a negative form. Take for example X to be an exponential distribution, $Pr(X \geq t) = e^{-\mu t}$, with some parameter μ which satisfies $\rho < 1$. For many queueing systems this particular choice of job-size distribution is easiest to analyze and leads to the cleanest results, thus it is, by far, the most popular job size distribution in queueing theoretic studies. In the mini-market problem with exponentially distributed job sizes, the answer to problem 1 can only be determined numerically and depends on ρ. The answer to problem 2 is that it does matter who works at the express counter. For example, numerically, it has been determined that when $\rho = 0.8$ and $c = 2$ it is better to assign the slower worker, Bob, to the express counter, [17].

For other job-size distributions, the opposite is true and the faster worker should handle the express counter. In addition, it is not clear that the result does not depend on ρ and c.

Regarding problem 3, there does not seem to be a nice formula relating an optimal cutoff to the optimal cutoff after the workers have switched counters.

2.3 A positive result

While the general situation seems bleak, we are happy to report that for some job-size distributions the answer to all three problems is simple and does not depend on the values of ρ and c. A mini-market manager who is lucky enough to observe one of these special job-size distributions will have a very easy time managing the mini-market.

We provide several examples. Define

$$q = q(t) = e^{-2\pi t}$$

1) Consider the Dedekind eta function

$$\eta(t) = q^{1/24} \prod_{n \geq 1} (1 - q^n)$$

The eta function plays an important role in number theory, for example, in the Kronecker limit formula and in the Chowla-Selberg formula, [12, 13, 37]. Let

$$\tilde{f}_\eta = \eta^8$$

2) We first consider some basic definitions related to characters. Let $(\mathbf{Z}/N)^*$ be the group of residues modulo N, which are prime to N, with multiplication as group operation. A **character** of $(\mathbf{Z}/N)^*$ is a homomorphism of groups

$$\chi = \chi_N : (\mathbf{Z}/N)^* \to \mathbf{C}^*$$

where $\mathbf{C}^*$ is the group of non zero complex numbers with multiplication as group operation. A character is **quadratic** if the image of χ is the subgroup $\{1, -1\}$. A character χ is **primitive** if it does not factor as

$$\chi = \chi_N : (\mathbf{Z}/N)^* \to (\mathbf{Z}/M)^* \to \mathbf{C}^*$$

where M divides N.

If χ is a primitive character, we say that N is the **conductor** of χ. A character χ is **odd** if $\chi(-1) = -1$ and **even** if $\chi(-1) = 1$. For each prime p there is a unique primitive quadratic character that takes the value 1 on m if the equation $x^2 = m \ Mod \ p$ is solvable, and -1

otherwise. This character is even if p has the form $4k+1$. We will denote this character by θ_p.

We can think of any character χ as a function on all integers if we define $\chi(n) = 0$ whenever, n and N have a nontrivial common divisor and by using the residue modulo N otherwise.

Assuming χ is a primitive, even, character, we define the **twisted θ** function

$$\theta_\chi(t) = \frac{1}{2}\sum_{n=-\infty}^{\infty} \chi(n)e^{-\pi n^2 t} = \sum_{n=1}^{\infty} \chi(n)e^{-\pi n^2 t} \tag{1}$$

note that if we take the trivial character $\chi = 1$ we get (after a slight modification) Jacobi's theta function

$$\theta_\chi(t) = \frac{1}{2}\sum_{n=-\infty}^{\infty} e^{-\pi n^2 t} = 1 + 2\sum_{n=1}^{\infty} e^{-\pi n^2 t} \tag{2}$$

The twisted theta functions play an important role in studying primes in arithmetic progressions. For a discussion related to probability theory see [6].

Assuming that p is a prime of the form $p = 4k+1$, we define

$$\tilde{f}_{\theta,p} = \theta^8_{\chi_p}$$

3) We introduce a family $X_{A,B}$ of job-size distributions, parametrized by a pair of integers $A, B \in \mathbf{Z}$. The integers A, B must satisfy some mild conditions. They must satisfy the inequality

$$\Delta_{A,B} = 4A^3 + 27B^2 \neq 0$$

In addition, for any prime p, if p^4 divides A then p^6 does not divide B.

Using A and B we define a sequence of integers a_n, $n \geq 1$. We let $a_1 = 1$. Given n, m with no common divisor we let $a_{nm} = a_n a_m$. By the unique decomposition of positive integers into a product of prime numbers we are left with the task of defining a_{p^k} for all prime powers p^k. Consider the equation

$$y^2 = x^3 + Ax + B \tag{3}$$

Let N_{p^k} denote the number of solutions to the equation (3), where we think of the integers A, B as numbers modulo p and x, y are in the field with p^k elements. Define $a_{p^k} = p^k - N_{p^k}$. Given the sequence a_n we define the power series $g_{A,B} = \sum_{n=1}^{\infty} a_n q^n$. Let $\tilde{f}_{A,B}(t) = g^2_{A,B}(t)$.

Theorem 1 *Let $\tilde{f} = \tilde{f}_\eta, \tilde{f}_{\theta,p}, \tilde{f}_{A,B}$ be as above. The following statements hold:*

0) The integral $\int_0^\infty \tilde{f}_{A,B} dt$ exists, thus we can define a density function

$$f(t) = \tilde{f}(t)/(\int_0^\infty \tilde{f} dt)$$

Consider the mini-market queueing system with job-size density function $f(t)$, and parameters ρ, c, then,

1) Let $c = 1$. There is an explicitly computable integer N, such that the optimal cutoff is given by $s_{opt} = 1/\sqrt{N}$, independent of ρ. When $f = f_\eta$ we have $N = 1$. When $f = f_{\theta,p}$ we have $N = p^2$. When $f = f_{A,B}$, N is divisible by the same primes as $\Delta_{A,B}$. If $p \neq 2, 3$ and p divides $\Delta_{A,B}$ which is the discriminant of the cubic $x^3 + Ax + B$, the cubic will have either one or two solutions in the algebraic closure of F_p, the field with p elements. If the cubic has two solutions we choose $N(p) = p$, if the cubic has a single solution we choose $N(p) = p^2$. If $p = 2, 3$, $N(p)$ has the form p^l for some $l < 10$, with the exact formula being described in [39]. We have $N = N_{A,B} = \prod_p N(p)$.

2) For any ρ and c it does not matter which assignment of workers to counters is chosen, namely,

$$E(W)_{opt}(f(t), \rho, c, \sigma) = E(W)_{opt}(f(t), \rho, c, \bar{\sigma})$$

3) The following formula relates the optimal cutoffs with respect to σ and $\bar{\sigma}$,

$$s_{opt}(f(t), \rho, c, \sigma) = \frac{1}{N s_{opt}(f(t), \rho, c, \bar{\sigma})}$$

independently of ρ and c.

As can be observed from the theorem, managing the mini-market with the job-size density functions as above, could not be easier. One observation that we wish to carry from this result is that functions which have a number theoretic flavor, may also have nice queueing theoretic properties. The other point is that when the job size distribution is related to number theory, queueing theoretic observables, in the present case, the optimal cutoff when $c = 1$ can have number theoretic significance.

The rest of the paper is devoted to explaining this theorem in its natural queueing theoretic and number theoretic contexts and how the two end up mixing together.

2.3.1 The Mellin transform

The moments of functions play an important role in both probability theory and number theory. In certain queueing systems the average waiting time is given in terms of the integer moments of the job size distribution. In number theory, taking moments allows one to pass between two families of interesting functions, the **automorphic forms** and the **L-functions**. The process of passing from a function to its moments is captured by the Mellin transform.

Definition 1 *Let $f(t)$ be a complex valued function defined on the non-negative reals* $(0,\infty)$. *For any complex number s, the* **Mellin transform** *of f evaluated at s is given by the formula*

$$L_f(s) = \int_0^\infty f(t)t^s \frac{dt}{t} \tag{4}$$

whenever it exists. Given an interval, $I \subset$ $(0,\infty)$, *we also define the* **incomplete Mellin transform** *with respect to I by the formula*

$$L_{f,I}(s) = \int_I f(t)t^s \frac{dt}{t} \tag{5}$$

By definition, the Mellin transform at s is the $s-1$ moment of f.

2.4 A family of involutions

A central figure in our story is a family of involutions on functions of $t > 0$.

Definition 2 *Let $f(t)$ be a function on $t > 0$. We define the* **Hecke involution with parameters** k, N *by the formula*

$$H_{k,N}(f)(t) = \hat{f}_{k,N}(t) = N^{-k/2}t^{-k}f(1/Nt) \tag{6}$$

The parameter k will be called the **weight** *of the involution $H_{k,N}$, while the parameter N will be called the* **level**.

If I is the interval $(a,b), [a,b), (a,b], [a,b] \subset (0,\infty)$ *then we define the* **level** N **dual of** I, *denoted $\hat{I}_N$, to be the image of I under the involution $t \to 1/Nt$.*

It is easy to check that $H_{k,N}$ is indeed an involution, i.e., when applied twice we obtain the identity map on functions. Obviously for any N, $I \to \hat{I}_N$ is an involution on intervals.

2.5 Relating moments and the involutions

The following basic computation (observation) of Riemann relates the incomplete Mellin transforms of f and of $\hat{f}_{k,N}$. While it is trivial, it is of central importance for our story.

Lemma 1 *Let k, N be some weight and level. Let f be a function on the positive reals and let $I \subset (0, \infty)$ be some interval, then, $L_{\hat{f},\hat{I}}(s)$ exists if and only if $L_{f,I}(k-s)$ exists and we have*

$$L_{\hat{f},\hat{I}}(s) = N^{k/2-s} L_{f,I}(k-s) \tag{7}$$

Proof: Letting $\tau = 1/Nt$, we have $\frac{d\tau}{\tau} = -\frac{dt}{t}$. Then,

$$L_{f,\hat{I}} = \int_{\hat{I}} \hat{f}(t) t^s \frac{dt}{t} = \int_{\hat{I}} N^{-k/2} t^{-k} f(1/Nt) t^s \frac{dt}{t}$$

$$= N^{-k/2} \int_I (\frac{1}{N\tau})^{s-k} f(\tau) \frac{d\tau}{\tau} = N^{k/2-s} L_{f,I}(s-k)$$

as required. *q.e.d.*

2.6 The Pollaczek-Khinchine formula and variants

Before dealing with the mini-market which has two servers, we consider a system with a single server (checkout counter). We assume Poisson arrivals, a job size distribution X and utilization ρ. We also assume a `FIFO` queue at the server. A basic result in queueing theory provides the average waiting time (not including service time) in such a queue in terms of X and ρ.

If we measure time in seconds rather than minutes, the average waiting time $E(W)$ will increase 60 fold. We would like an invariant performance measure. Therefore, we introduce the average normalized waiting time

$$E(\tilde{W}) = E(W)/E(X) \tag{8}$$

Since a change in time units will affect proportionately both $E(W)$ and $E(X)$ we get an invariant performance measure.

Assume that f is the density function for the job size distribution X and that L_f is the Mellin transform of f. The Pollaczek-Khinchine formula which was proved independently by F. Pollaczek, [31], and A. Khinchine, [24] computes $E(\tilde{W})$ in terms of the moments of X. We will use the Mellin transform notation to express the result.

$$E(\tilde{W}) = \frac{\rho}{2(1-\rho)} \frac{L_f(1)L_f(3)}{L_f(2)^2} \tag{9}$$

It may seem odd that we have included the term $L_f(1) = \int_0^\infty f(t)dt$, since it equals 1 for any density function. However, we note that it converts $E(\tilde{W})$ into an expression which is invariant if we compute using cf rather than f, for any constant $c > 0$. This will turn out to be convenient since the involutions $H_{N,k}$ do not preserve the integral condition for densities.

The requirement that a formula like that for $E(\tilde{W})$ will be invariant under multiplication by a constant translates into the condition that the number of Mellin transforms in the numerator and denominator should be equal. The condition that the formula should be invariant under change of time units translates into the condition that the sum of values at which the Mellin transform is evaluated is equal in the numerator and the denominator.

We can apply a similar procedure to other popular target functions.

Definition 3 *Given a job which waited time w and whose size is x we define the* **slowdown** *as w/x.*

Average slowdown is a popular measure of a system's performance. In a `FIFO` queue, there is no relation between the waiting time of a job and its size, so $E(S) = E(W)E(X^{-1})$, where S denotes the slowdown. This leads to the formula

$$E(S) = \frac{\rho}{2(1-\rho)} \frac{L_f(0)L_f(3)}{L_f(2)} \tag{10}$$

which after our homogenization procedure leads to

$$E(S) = \frac{\rho}{2(1-\rho)} \frac{L_f(0)L_f(3)}{L_f(1)L_f(2)} \tag{11}$$

There are also formulas for higher moments of the waiting time W. As before $E(W^k)$ is not a time unit invariant quantity. To normalize we consider the random variable $\tilde{W}_m = \frac{W^m}{E(X^m)}$. For $E(\tilde{W}_m)$, the normalized m'th moment of waiting time, we have a recursive formula due to Takacs, [40]

$$E(\tilde{W}_m) = \frac{\rho}{1-\rho} \frac{1}{E(X)E(X^m)} \sum_{i=1}^{m} \frac{B_{m,i}}{i+1} E(X^{i+1})E(X^{m-i})E(\tilde{W}_{m-i}) \tag{12}$$

where $B_{m,i}$ is the binomial coefficient and $E(\tilde{W}_0)$ is 1 by convention. We can see by induction that the expressions for $E(\tilde{W}_m)$ will be invariant both for multiplication of the density by a constant and for a change of time units. We also see from this formula that all moments of waiting time are rational combinations of the Mellin transform evaluated at integer points and the utilization. The first summand of $E(\tilde{W}_m)$ is

$$\frac{\rho}{(m+1)(1-\rho)} \frac{L_f(1)L_f(m+2)}{L_f(2)L_f(m+1)}$$

Instead of servicing customers in the order in which they enter the queue, we may consider other orderings such as `LIFO` (Last In First Out), where the latest arrival to the queue is served first. Another option is to service a random customer from the queue. It can be shown that these queue management methods lead to the same value of $E(\tilde{W})$, but have different formulas for $E(\tilde{W}_m)$. In any case, all moments of waiting time are again combinations of the utilization and the Mellin transform evaluated at integer points. The first summand of $E(\tilde{W}_m)$ in the case of `LIFO` is

$$\frac{\rho}{(m+1)(1-\rho)^{m+1}}\frac{L_f(1)L_f(m+2)}{L_f(2)L_f(m+1)}$$

If we know the sizes of the jobs in the queue we can also select the next customer to be serviced, based on the job size. For example, it seems to make sense to allow customers with small jobs to pass customers with large jobs, this policy is called `SJF`, Shortest Job First. We can similarly define the somewhat less intuitive `LJF` Longest Job first. In policies such as `SJF` or `LJF` each job size has its own waiting time since jobs of different sizes have different priorities. Let $\tilde{W}_{f,SJF}(t) = E(W_t)/E(X)$ be the average normalized waiting time of a job of size t, when the policy is `SJF` and the job size density is f and similarly $\tilde{W}_{f,LJF}(t)$ the corresponding quantity in a `LJF` managed queue. Using well know formulas, see [26], we can express $\tilde{W}_{f,SJF}$ in terms of incomplete Mellin transforms as

$$N_{f,SJF}(t) = \frac{\rho}{2(1-\rho\frac{L_{f,I_t}(2)}{L_f(2)})^2}\frac{L_f(1)L_f(3)}{(L_f(2))^2} \tag{13}$$

and similarly for `LJF` we have

$$N_{f,LJF}(t) = \frac{\rho}{2(1-\rho\frac{L_{f,I^t}(2)}{L_f(2)})^2}\frac{L_f(1)L_f(3)}{(L_f(2))^2} \tag{14}$$

where $I_t = (0,t)$ and $I^t = (t,\infty)$.

So far we have considered only Poisson arrivals. More generally, one can consider an arrival process where the times between successive arrivals of customers are i.i.d. random variables with distribution Y which is not necessarily exponential. Let g denote the density function associated with Y, assuming it exists. In general there is no nice expression for the average waiting time in such a queue, however, when $\rho \to 1$, a situation known as a heavy traffic limit, it has been shown by Kingman, [25], that the normalized average waiting time $E(\tilde{W})$ is asymptotically given by,

$$E(\tilde{W}) \sim \frac{1}{1-\rho}[\frac{\rho L_f(1)L_f(3)}{L_f^2(2)} + \frac{L_g(1)L_g(3)}{\rho L_g^2(2)} - (\rho + \frac{1}{\rho})]$$

where $\sim$ means that the ratio of the two sides approaches 1 as $\rho \to 1$.

One important consequence of the P-K formula is that large variance (second moment) in the job size distribution leads to large waiting times. This motivates the notion of "express lines" with cutoffs as we have encountered in the mini-market problem. In such queueing systems each host is responsible only for jobs in a certain range. This reduces the variance at the host. We now introduce SITA policies, which formalize the notion of express lines. These policies have been introduced in the computer science setting in [21].

Definition 4 *A* SITA *policy can be described as follows:*
*Consider a system with h hosts, numbered $1, \ldots, h$ and with speeds c_i with $c_1 = 1$. Let $s_1 < s_2 < \ldots < s_{h-1}$ be a set of processing time cutoffs with respect to the first host. We let $s_0 = 0$ and $s_h = \infty$. Let σ be a permutation on the set of hosts $1, ..., h$. We assume that job-sizes are known. An incoming job of size s with respect to the first host is dispatched to host $\sigma(i)$, such that $s_{i-1} \leq s < s_i$. Requests are serviced in each host in a first come, first served (*FIFO*) order.*

Assume that the arrivals to the SITA system are Poisson. We let X be the job-size distribution with respect to the first host. Let us extend the P-K formula to cover SITA systems with a given set of cutoffs $s_1 < s_2 < \ldots < s_{h-1}$. Let $I = (a, b]$ be some interval with $a, b \geq 0$. We let X_I be the distribution of job-sizes restricted to the interval I. If X is given by some density f, then the restriction of f to the interval I is, up to a constant multiple, the density for X_I. In terms of the incomplete Mellin transform $L_{f,I}(s) = \int_I f(t)t^{s-1}dt$ we have

$$E(X_I^k) = \frac{L_{f,I}(k+1)}{L_{f,I}(1)} \tag{15}$$

Let $I_j = (s_{j-1}, s_j]$. Let $E(W_j)$ be the average waiting time (service not included) of jobs in the j'th queue of a SITA system. If f is a proper density, namely, $L_f(1) = 1$, then the portion of jobs in the j'th queue is $L_{f,I_j}(1)$.

We conclude that $E(W) = \sum_j L_{f,I_j}(1)E(W_j)$. We can compute each individual $E(W_j)$ using the P-K formula for a single queue in terms of incomplete Mellin transforms.

$$\begin{aligned} E(W_j) &= \frac{\rho}{c_{\sigma(j)}} \frac{L_{f,I_j}(1)}{L_f(2)} \frac{L_{f,I_j}(3)}{L_{f,I_j}(1)} (2(1 - \frac{\rho}{c_{\sigma(j)}} \frac{L_{f,I_j}(2)}{L_f(2)}))^{-1} \\ &= \frac{\rho}{2c_{\sigma(j)}} (1 - \frac{\rho}{c_{\sigma(j)}} \frac{L_{f,I_j}(2)}{L_f(2)}))^{-1} \frac{L_{f,I_j}(3)}{L_f(2)} \end{aligned}$$

To get the weighted contribution of server j, to the average normalized waiting time $E(\tilde{W})$, we need to multiply by $L_{f,I_j}(1)$, the portion of jobs arriving at host j, and to divide by $L_f(2) = E(X)$. Calling the weighted contribution of host j to $E(\tilde{W})$ by $E_j(\tilde{W})$ we have

$$E_j(\tilde{W}) = \frac{\rho}{2c_{\sigma(j)}}(1 - \frac{\rho}{c_{\sigma(j)}}\frac{L_{f,I_j}(2)}{L_f(2)})^{-1}\frac{L_{f,I_j}(1)L_{f,I_j}(3)}{L_f(2)^2} \tag{16}$$

and

$$E(\tilde{W}) = \sum_j E_j(\tilde{W}) \tag{17}$$

Starting with (10) we can also develop the analogous formula for the contribution of the j'th host to average slowdown in a `SITA` queue

$$E_j(S) = \frac{\rho}{2c_{\sigma(j)}}(1 - \frac{\rho}{c_{\sigma(j)}}\frac{L_{f,I_j}(2)}{L_f(2)})^{-1}\frac{L_{f,I_j}(0)L_{f,I_j}(3)}{L_{f,I_j}(1)L_{f,I_j}(2)}. \tag{18}$$

3 Duality theory

We are in a position to bring together the analysis via the P-K formula of `SITA`, `SJF` and `LJF` queues with Riemann's lemma on the behavior of the Mellin transform w.r.t. the involutions $H_{k,N}$. The basic result is the following, see [17] and [1].

Theorem 2 *Let f be a density for which $L_f(2), L_f(3) < \infty$. For a single server queue with Poisson arrivals, fixed utilization ρ and job size density f, we have, for any level $N > 0$,*

$$E(\tilde{W})(f,\rho) = E(\tilde{W})(H_{4,N}(f)) \tag{19}$$

In addition,

$$H_{0,N}(\tilde{W}_{f,SJF}) = \tilde{W}_{H_{4,N}(f),LJF} \tag{20}$$

We also have

$$E(S)(f) = E(S)(H_{3,N}(f)) \tag{21}$$

Let $E(\tilde{W})(g,f,\rho)$ denote the normalized waiting time of a single server queue with i.i.d. inter-arrival times with density g, i.i.d. service times with density f and utilization ρ. Then, for any level N,

$$lim_{\rho\to 1}\frac{E(\tilde{W})(f,g,\rho)}{E(\tilde{W})(H_{4,N}(f),H_{4,N}(g),\rho)} = 1$$

Consider a `SITA` queue with h linearly coupled hosts with speeds given by the vector $C = (c_1,...,c_h)$, cutoffs $s_1,...,s_{h-1}$, utilization ρ, job size density function f and a permutation σ, where host $\sigma(i)$ is assigned to handle jobs whose size s is in the interval $[s_{i-1}, s_i)$.

Let $E(N)(f, h, \rho, C, \sigma, s_1, \ldots, s_{h-1})$ denote the average normalized waiting time of the `SITA` *queue*

Let $\hat{s} = \frac{1}{Ns}$. Then, the following dualities hold for any level $N > 0$,

$$E(\tilde{W})(f, h, \rho, C, \sigma, s_1, \ldots, s_{h-1}) = E(\tilde{W})(H_{4,N}(f), h, \rho, C, \bar{\sigma}, \hat{s}_{h-1}, \ldots, \hat{s}_1) \tag{22}$$

Proof: The single host results follow from Riemann's lemma and the P-K average normalized waiting time expressions for the various queues. For `SITA` queues, let $I_j = (s_{j-1}, s_j]$. Then, in the notation of lemma 1, $\hat{I}_j = (\hat{s}_{h-j}, \hat{s}_{h-j+1}]$. The host which is assigned to handle the j'th interval of the original queue, handles the $h-j-1$ interval in the dual system, i.e., it handles the dual interval. Denoting the weighted contribution to normalized average waiting time of the j'th host to the original queue by $E_j(\tilde{W})$ and to the dual queue by $E_j(\hat{\tilde{W}})$, we have by equation (16) and lemma 1 that $E_j(\tilde{W}) = E_{h-j+1}(\hat{\tilde{W}})$, which proves the assertion for `SITA` systems. *q.e.d.*

We consider the mini-market problem using duality. We need a result of Harchol-Balter and Vesilo, [22], that the optimal cutoff in a `SITA` system with two identical hosts is essentially unique.

Theorem 3 *Let f be a density function for which $L_{f,I_t}(2)$ is a strictly increasing function of t. A* `SITA` *queue with job size density $f > 0$, load ρ, and two identical hosts has a unique cutoff, s, which minimizes average normalized waiting time.*

Proof: Recall that $I_t = (0, t]$. Let $m_i(t) = L_{f,I_t}(i+1)$ be the i'th incomplete moment of f. Using the assumption that $f > 0$ we can re-parametrize the time coordinate via $m_1(t)$. From the definition we see that $dm_i(t)/dt = t^i f(t)$ and so $dm_i/dm_1 = (dm_i/dt)/(dm_1/dt) = t^{i-1}$. We consider the contribution of the first host to average waiting time. In terms of the m_i, it is given by $E_1(\tilde{W}) = \frac{1}{2(1-\rho\frac{m_i}{m_i^f})}\frac{m_0 m_2}{(m_1^f)^2}$ where $m_1^f = \int_0^\infty t f(\tau) d\tau$. We will show that as a function of m_1 this is a convex function. We need to show that $\frac{d^2}{d^2 m_1} E_1(\tilde{W}) \geq 0$. After some tedious manipulations it boils down to showing that $\frac{d^2}{d^2 m_1} m_0 m_2 \geq 0$. We have $\frac{d}{dm_1} m_0 m_2 = \frac{m_2}{t} + t m_0$. Differentiating again and using $dt/dm_1 = 1/(dm_1/dt) = \frac{1}{tf(t)}$ we get

$$\frac{d^2}{d^2 m_1} m_0 m_2 = 2 + (m_0 - \frac{m_2}{t^2})\frac{1}{tf(t)}$$

but, $m_0(t)t^2 = \int_0^t t^2 f(\tau) d\tau > 0 \int_0^t \tau^2 f(\tau) d\tau = m_2(t)$ which implies that the second summand is non-negative and therefore that the contribution of the first host is convex. For the contribution of the second host $E_2(\tilde{W})$ we can use duality to show that it equals the contribution of the first host in the queue with the dual job size density $H_{4,s}(f)$. The parameter m_1 is replaced in this case by $m_1^f - m_1$. However, this does not change the second derivative so we conclude that this contribution is also convex and therefore $E(\tilde{W})$ is convex as a function of m_1. Since a convex function with a strictly positive second derivative has a unique minimum over an interval we obtain the desired result. *q.e.d.*

Definition 5 *We say that a density f is* **self-dual of weight** *k if for some level $N > 0$ we have $f = H_{k,N}(f)$. We say that k is a weight of f and N is a level or conductor of f. We will denote the conductor by $N_f = N_{f,k}$, when the weight is explicit.*

The following result shows how to solve the mini-market management problem for self-dual job-size densities of weight 4.

Theorem 4 *If $k = 4$ is a weight of f, the job-size density of a mini-market queue and $L_f(3) = E(X^2) < \infty$, then:*

1) The optimal cutoff s_{opt} for minimizing average waiting time when $c = 1$ is

$$s_{opt} = 1/\sqrt{N_f} \tag{23}$$

This is also the queue length balanced solution.

2) The assignment of servers to counters is not important

$$E(\tilde{W})_{opt}(f, \rho, c, \sigma) = E(\tilde{W})_{opt}(f, \rho, c, \bar{\sigma})$$

3) We have

$$s_{opt}(f, \rho, c, \sigma) = \frac{1}{N_f s_{opt}(f, \rho, c, \bar{\sigma})}$$

Proof: From theorem 2 it is clear that if f is self-dual of weight 4 and the hosts are equally powerful, then, if s_{opt} is an optimal cutoff, so is $1/(N_f s_{opt})$. By uniqueness we have that the two values are equal and we obtain the first part of the theorem. The second part and third part also follow directly from theorem 2. *q.e.d.*

4 Self dual densities

From theorem 4 we see that all that is left to show, in order to prove theorem 1, is that the functions $f_\eta, f_{\theta,\chi}, f_{A,B}$ are self-dual of weight 4.

4.1 Self dual densities in queueing theory

Before we consider these functions f that come from number theory, we consider self dual distributions which often appear in the queueing theory literature when modeling job-size distributions. It will be more convenient to consider general non-negative functions f on the positive reals. In such cases it should be understood that actual densities are obtained from f by restricting to any interval I for which $0 < \int_I f(t)dt < \infty$ and normalizing. In addition, following our discussion so far, we will not distinguish between the functions f and cf when $c > 0$. We will refer to non-negative functions as **laws**.

4.1.1 Pareto distributions

Definition 6 *A* **Pareto law** *with parameter α is a function f of the form $f_\alpha(t) = t^{-\alpha-1}$. If the function is restricted to a finite interval $I = [a, b]$, $0 < a < b < \infty$, we say that it is Bounded Pareto and denote it by $f_{\alpha,I}$.*

Pareto distributions also have a discrete counterpart which has non-negative values which is known as **Zipf's law**.

Some reasons why Pareto and Zipf laws should be commonly observed is given in Terence Tao's blog entry, [41], where it is explained that these distributions have some universality properties as limits, much in the same way that the normal distribution is universal for limits of sums of i.i.d. random variables.

Pareto distributions have a very important property, they are scale invariant. Assume $c > 0$. For an interval $I = [a, b]$ denote by cI the scaled interval $[ca, cb]$, which may be obtained from I by scaling the time unit. For a function f, let f_I denote its restriction to I. Up to a multiplicative constant we have $f_{\alpha,I}(ct) = f_{\alpha,cI}(t)$, which means that the Pareto law looks the same at all scales. A short computation shows that

$$H_{k,N}(f_\alpha) = f_{k-\alpha-2} \tag{24}$$

In particular $f_{\alpha,l}$ is self-dual of weight $(k/2) - 1$ and any N. As a family Pareto laws are closed under $H_{k,N}$ for all N and k.

For $k = 4$, which is associated by theorem 2 with normalized average waiting time, we get that $f_{1,l}$ is self-dual and more generally $f_{\alpha,l}$ and $f_{2-\alpha,l}$ are dual. Similarly, for $k = 3$, which is associated with average slowdown we get that $f_{1/2,l}$ is self-dual and more generally $f_{\alpha,l}$ and $f_{1-\alpha,l}$ are dual.

4.1.2 Log-normal laws

Definition 7 *A* **log-normal** *law is a function of the form*

$$f(t) = \frac{1}{t}Exp(\frac{-(\log(t)-\mu)^2}{2V}) \tag{25}$$

where $V, \mu > 0$ are parameters.

The log-normal distributions are obtained when we consider the case $V > 0$. The normalizing factor for the denisty is $\frac{1}{\sqrt{2\pi V}}$. As the name suggests, the density of the distribution is obtained by plugging $\log(t)$ as the variable in the formula for the density of the normal distribution with average μ and variance V. The $\frac{1}{t}$ term comes from $d\log(t) = \frac{dt}{t}$. Log-normal distributions are commonly used for modeling job size distributions, [28, 29, 33, 30].

From the point of view of duality log-normal laws have a remarkable property. For any k there exists an $N(k)$ such that f is self dual w.r.t. $H_{k,N(k)}$. This has been observed in [17]. We consider more generally the family of functions of the form

$$f(t) = Exp[a\log^2(t) + b\log(t)] \tag{26}$$

which up to a constant multiple contain all products of a Pareto law and a log-normal law. We call these functions **generalized log-nornal** functions After a short calculation we get that

$$f(t) = ct^{-2a\log(N)+2b}f(1/Nt)$$

for some constant c, hence for any level N, the function is self-dual w.r.t. $H_{k,N}$ where $k = 2a\log(N) - 2b$. We claim that this property essentially characterizes these functions.

Theorem 5 *Let g be a measurable function such that for any $N > 0$ there exists $k = k(N)$ such that $g = H_{k(N),N}(g)$, then g is a constant multiple of a generalized log-normal function.*

Proof: It is somewhat easier to follow the argument if we make the change of variable $x = \log(t)$ and consider g as a function of x. The mapping $t \rightarrow \frac{1}{Nt}$ becomes $x \rightarrow -\log(N) - x$. Recall that the reflection of the real numbers centered at a point d is given by $x \rightarrow 2d - x$, hence denoting this map by $I_d(x)$ and setting $d = -\frac{1}{2}\log(N)$ we see that the relation $g(t) = N^{-k/2}t^{-k}g(\frac{1}{Nt})$ becomes

$$g(x) = e^{dk}e^{-xk}g(I_d(x)) \tag{27}$$

Consider $N_1 = 1$, $N_2 = 1/e$, the corresponding values of d are $d_1 = 0$, and $d_2 = 1/2$. Assume that the weights of g w.r.t. these two values of N are k_1, k_2 respectively. We can choose a generalized log-normal density which will have these weights, namely, with $b = -k_1/2$ and $a = -1/2(k_2 - k_1)$. Taking the ratio of g with this generalized log-normal we obtain a non vanishing function h which has weight $k_1 = k_2 = 0$ w.r.t. the levels 1 and $1/e$ respectively, or after the change of variable $h(x) = h(I_0(x)) = h(I_{1/2}(x))$. If we have two reflections I_{d_1} and I_{d_2} then their composition is $x \to 2d_2 - (2d_1 - x) = 2(d_2 - d_1) + x$, hence we conclude that h has period 1. Consider the reflection $I_{1/4}$ which corresponds to $N = e^{-1/2}$. By our assumption on g and the properties of generalized log-normal functions we know that h will satisfy (27) for $d = 1/4$ with some value of k, say $k(1/4)$. Taking $x = -3/4$ we see that $h(-3/4) = e^{k(1/4)}h(5/4) = h(-3/4)$ from which we conclude that $k(1/4) = 0$ and hence that h has period $1/2$ by taking $d_1 = 0$ and $d_2 = 1/4$. Repeating with $d = 1/8$ and so on we see that h has period 2^{-l} for all integers $l \geq 0$. However, using exercises 7.3 and 7.4 in [34] it is easy to show that a measurable function which has arbitrarily small periods is constant almost everywhere. Let c be the constant a.e. value of h. For any given x we can find an a such that $h(a) = h(a + x) = c$, hence h has weight 0, i.e., is periodic, for any x and therefore is constant everywhere. We conclude that g is a generalized log-normal function as desired. *q.e.d.*

4.2 Self dual densities in number theory

Having considered a few examples which are commonly considered in queueing theory, we turn to the examples from number theory.

Let H be the complex upper half plane consisting of complex numbers $z = x + yi$ such that $y > 0$. Letting $q = e^{2\pi iz}$, we can think of η as a function on H. We obtain the previous definition if we set $z = z(t) = it$, $t > 0$. It can be shown, see [15] for a proof, that

$$\eta(-1/z) = \sqrt{z/i}\eta(z)$$

For $\tilde{f}_\eta$ this means that

$$\tilde{f}_\eta = H_{4,1}\tilde{f}_\eta$$

we turn to $\tilde{f}_{\theta,p}$. Given a primitive character χ of conductor N, we may define the **Gauss sum**

$$\tau_\chi = \sum_{\mathbf{Z}_N} \chi(n)e^{2\pi in/N}$$

It can be shown that $|\tau_\chi|^2 = N \neq 0$. It can also be checked that $\tau(\bar{\chi}) = \chi(-1)\overline{\tau(\chi)}$ from which we conclude that when χ_p is even, we have $\tau(\chi_p) = p^{1/2}$ or $\tau(\chi_p) = -p^{1/2}$. Gauss showed that actually $\tau_{\chi_p} = p^{1/2}$. The following formula, see [11], can also be verified

$$\chi(n) = \frac{\chi(-1)\tau_\chi}{N} \sum_{m \in (\mathbf{Z}_N)^*} \bar{\chi}(m) e^{2\pi i n m/N} \tag{28}$$

A famous result of Jacobi is that his theta function satisfies the functional equation

$$\theta(t) = t^{-1/2}\theta(1/t)$$

The result can be proved using the Poisson summation formula for Fourier coefficients. Repeating the proof in conjunction with equation (28) and Gauss' computation of $\tau(\chi_p)$ leads to the twisted functional equation for these characters

$$\theta_{\chi_p}(t) = p^{-1/2} t^{-1/2} \theta_\chi(\frac{1}{p^2 t}) \tag{29}$$

see [11] for details. the equations above show that as generalized densities $\theta_{\chi_p} = H_{1/2,N^2}(\theta_{chi_p})$ or

$$\tilde{f}_{\theta,p} = H_{4,N^2}(\tilde{f}_{\theta,p})$$

To discuss the last example, $\tilde{f}_{A,B}$, we need to consider modular forms. The group $GL_2(\mathbf{R})^+$ of real 2 by 2 matrices with positive determinant acts on upper half plane H. The action of $\gamma \in GL_2(\mathbf{R})^+$ is given by

$$\gamma(z) = (az+b)/(cz+d) \tag{30}$$

where the matrix γ is given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{31}$$

It is easy to check that for such matrices γ we have

$$Im(\gamma(z)) = det(\gamma)\frac{Im(z)}{|cz+d|^2} \tag{32}$$

which also verifies that H is indeed preserved by this mapping. In addition it is easy to check that this is a group action, namely

$$\gamma_1(\gamma_2(z)) = (\gamma_1\gamma_2)(z) \tag{33}$$

Definition 8 *We define the* **weight** k **action** *of* $\gamma \in GL_2(\mathbf{R})^+$ *on a function* f *defined on* H *by the formula*

$$f_{\gamma,k}(z) = det(\gamma)^{k-1}(cz+d)^{-k} f(\gamma(z)) \tag{34}$$

It is easy to verify using (33) that this is also a group action

$$f_{\gamma_1\gamma_2,k} = (f_{\gamma_2,k})_{\gamma_1,k} \tag{35}$$

Definition 9 *The* **congruence subgroup** $\Gamma_0(N) \subset SL_2(\mathbf{Z})$ *consists of those matrices with* $c = 0$ $(Mod\ N)$*.*

Definition 10 *We say that a holomorphic function on* H *is a* **modular form of weight** k **and level** N **for** $\Gamma_0(N)$ *if it satisfies the weight* k *modular functional equation*

$$f_{\gamma,k} = f \tag{36}$$

for all $\gamma \in \Gamma_0(N)$*. We denote the set of modular forms by* $M_k(N)$*. If in addition to being a modular form, for any* $\delta \in SL_2(\mathbf{Z})$*, the function* $f_{\delta,k}(z)$ *decreases exponentially fast as* $Im(z) \to \infty$*, we say that* f *is a* **cusp form**. *We denote the set of cusp forms of weight k and level N by* $S_k(N)$*.*

Since $\Gamma_0(N)$ contains the matrix

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{37}$$

the modularity condition (36) states that $f(z) = f(z+1)$, which means that f will have a Fourier series expansion in the variable $q = e^{2\pi i z}$. The exponential decay implies that for a cusp form the expansion can be written as $f(z) = \sum_{n=1}^{\infty} a_n q^n$.

The involution sending $f(z)$ to

$$\hat{f}_{k,N}(z) = N^{-k/2}(z/i)^{-k} f(-1/Nz) \tag{38}$$

maps $S_k(N)$ to itself. To see why, consider the element $w_N \in GL_2(\mathbf{Q})^+$

$$\begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix} \tag{39}$$

The action of w_N on H is given by $w_N(z) = \frac{-1}{Nz}$ and so, $f_{\omega_N,k} = c_{k,N}\hat{f}_{k,N}$, where $c_{k,N}$ is a constant. It is easy to verify that $w_N^{-1}\Gamma_0(N)w_N = \Gamma_0(N)$ and from this we can deduce that if $f \in S_k(N)$ then so is $f_{w_N,k}$ and consequently $\hat{f}_{k,N} \in S_k(N)$ as well.

The relation between $\hat{f}_{k,N}$ and the queueing theoretic duality map is given by restricting to the upper imaginary axis $z = it$. Comparing equations (6) and (38), and using the mapping

$g(t) = f(it)$ to move between functions with a complex variable and functions with a positive real variable, we see that the involutions $H_{k,N}$ and $\hat{f}_{k,N}$ coincide when $z = it$.

We observe that if K, N are integers such that K divides N, then $\Gamma_0(N) \subset \Gamma_0(K)$. Therefore, a modular form w.r.t. $\Gamma_0(K)$ is also a modular form w.r.t. $\Gamma_0(N)$. In addition, if $N = KD$, consider the matrix δ_D given by

$$\begin{pmatrix} D & 0 \\ 0 & 1 \end{pmatrix} \tag{40}$$

The matrix δ_D induces on the upper half plane H the mapping $\delta_D(z) = Dz$. It is easy to verify that

$$\Gamma_0(N) \subset \delta_D^{-1}\Gamma_0(L)\delta_D$$

hence for any $f \in S_k(L)$ we have

$$f_{\delta_D,k} = D^{k-1}f(Dz) \in S_k(N)$$

Definition 11 *A form $f \in S_k(N)$ is said to be an* **old form** *if it is in the subspace of $S_k(N)$ which is generated by elements of $S^k(K)$ for K dividing N, $K \neq N$, and by elements of the form $f((N/K)z)$ for $f \in S_k(K)$.*

4.2.1 Hecke operators

There is a family of natural operators T_n, known as the Hecke operators, acting on $S_0(N)$. This family is of central importance in number theoretic applications of modular forms.

in terms of the q expansion we have

$$T_n(f) = \sum_m [\sum_{d|gcd(m,n)} d^{k-1}a_{mn/d^2}(f)]q^m \tag{41}$$

We have the multiplicative relation

$$T_{nm} = T_n T_m \tag{42}$$

whenever n, m are prime to each other, namely, $gcd(n, m) = 1$.

For p not dividing N, we have the recursive formula

$$T_{p^r} = T_p T_{p^{r-1}} - p^{k-1}T_{p^{r-2}} \tag{43}$$

for $r \geq 2$, where T_1 is the identity. For $p|N$ we get

$$T_{p^l} = T_p^l \tag{44}$$

Definition 12 *A cusp form which is an eigenvector for the operators T_n for all n, which is not an old form and which is normalized so that $a_1 = 1$ is called a* **Hecke newform***.*

Taking $m = 1$ in equation (41) we see that for a Hecke newform, $f = \sum_{n\geq 1} a_n q^n$, the coefficient a_n is the eigenvalue corresponding to the operator T_n. The following result, see [15], for a proof shows that hecke newforms are also eigenvectors for $H_{k,N}$.

Theorem 6 *Let f be a Hecke newform, and let $g(t) = f(it)$, then either, $g = H_{k,N}(g)$ or $g = -H_{k,N}(g)$.*

Definition 13 *Let $f \in S_k(N)$ and let $f = \sum_n f_n q^n$ be its Fourier expansion. Let χ be a primitive Dirichlet character modulo D, with g.c.d.$(D, N) = 1$. We define the* **twisting of** *f by the character χ to be the function f^χ with expansion*

$$f^\chi = \sum_n \chi(n) f_n q^n$$

Using (28) it can be shown that restricting to the positive imaginary line

$$H_{k,ND^2}(f^\chi) = \chi(N)\frac{\tau_\chi^2}{D} H_{k,N}(f)^{\bar{\chi}} \tag{45}$$

From the above discussion, we conclude that all the twists of a form in $S_k(N)$ are related via the involutions $H_{k,M}$ for various M to all the twists of the dual form. We may ask if in analogy with the case of log-normal distributions having these many functional duality relations characterizes such forms. We can now state Weil's converse theorem, [43], which answers this question positively.

Theorem 7 *Let $f = \sum_n f_n q^n$ and $g = \sum_n g_n q^n$, where f_n, g_n are some sequences of polynomial growth, namely, $|f_n|, |g_n| \leq cn^r$, for some $c, r > 0$. If for any primitive Dirichlet character χ of conductor D, which satisfies g.c.d.$(D, N) = 1$, the functions $L_{f,\chi}$ and $L_{g,\bar{\chi}}$ can be analytically continued to the entire plane, are bounded in every vertical strip $A \leq Re(s) \leq B$ and satisfy the functional equation (45), then, $f \in M_k(N)$.*

4.2.2 Some examples of modular forms

From the functional equation $\eta(-1/z) = \sqrt{z/i}\eta(z)$ it follows that $\tilde{\Delta} = \eta^{24}$ satisfies $\eta(-1/z) = z^{12}\eta(z)$ and $\eta(z+1) = \eta(z)$. In addition, it has a q expansion with $a_0 = 0$, hence it is in $S_{12}(1)$. A dimension calculation shows that $dim(S_{12}(1)) = 1$, hence $\tilde{\Delta}$ is a Hecke newform. The function $\Delta = (2\pi)^{12}\tilde{\Delta}$ is known as the discriminant and essentially coincides with the function Δ we presented above. We note that η does not vanish on the upper half plane and hence, the same holds for Δ. Via a dimension computation and the non vanishing of Δ it can be further shown, see [15], that if $k(N+1) = 24$ and k is even, then $(\eta(z)\eta(Nz))^k \in S_k(N)$ and that these spaces are 1 dimensional. We conclude that these functions are Hecke newforms which are positive on the upper imaginary axis. The following result is known as the Taniyama-Shimura conjecture. It was established in [10], relying heavily on the work of [44] and [42], which established an important special case. The proof is extremely complicated and is considered as one of the highlights of mathematics. For an overview we refer the reader to [14].

Theorem 8 *The functions $g_{A,B}$ are Hecke newforms in $S_2(N_{A,B})$.*

From this theorem and theorem 6 we conclude that $\tilde{f}_{A,B} = g^2_{A,B}$ satisfies

$$\tilde{f}_{A,B} = H_{4,N_{A,B}}(\tilde{f}_{A,B})$$

Proof of theorem 1: The mini-market theorem follows from the functional equations that the functions $\tilde{f}$ satisfy and theorem 4 once we know claim 0 of the theorem. However, the functional equation of $\tilde{f}$ and the exponential decay of $\tilde{f}$ at infinity implies exponential decay near $t = 0$ as well, hence $\int_0^\infty \tilde{f}_{A,B} dt$ exists. *q.e.d.*

Remark: One major motivation for proving that $g_{A,B}$ is a Hecke newform was proving the analytical continuation and functional equation for the function $L_{g_{A,B}}$, as was the case with Riemann's zeta function, using the Jacobi theta function. However, another major motivation (in fact, the main motivation for Wiles) is that together with results of Frey, Serre and Ribet it resolved Fermat's last theorem, See [14] for details.

5 Positivity

We can ask whether $\theta_p = \theta_{\chi_p}$ is a positive function. This question is very old and has been studied extensively. It was proved in [2] that for any $k > 0$, for almost all primes f_p has more than k zeroes in the range $0 < x < 1$, hence having positive θ_p is unfortunately an asymptotically rare phenomenon. It was asked in [20] and again in [2] whether the number of positive f_p is finite. This remains an open problem. For the same reasons that we inquired about positivity of the θ_p we can ask whether Hecke newforms, and in particular the functions $g_{A,B}$ are positive. The moments of these functions are conjecturally of great arithmetic interest and hence also the moments of waiting time in an M/G/1 queue with such densities, see [4, 5, 8, 9].

As an example of the arithmetic interest in moments, consider the basic step of converting $g = g_E = g_{A,B}$ from a generalized density into an actual probability density. The function g_E was normalized in such a way that $a_1 = 1$. This makes sense from a number theory point of view since it made the coefficient a_n coincide with the eigenvalue of the Hecke operator T_n. However, for the purposes of probability theory we need to divide g_E by the constant $\int_0^\infty g_E(t)dt$.

By definition, this constant is $L_g(1)$. The arithmetic interpretation of this normalizing term is the subject of the Birch and Swinnerton-Dyer conjecture (BSD), [7]. Obviously, if g is strictly positive then $L_g(1) > 0$ will be positive. By a deep result of Kolyvagin [27], combined with the Taniyama-Shimura conjecture, the condition $L_g(1) \neq 0$, is equivalent to having only a finite number of rational solutions to the equation $y^2 = x^3 + Ax + B$, providing an arithmetic necessary condition for positivity of g. This is a necessary condition, but is not sufficient.

We can test numerically whether functions of the form $g_{A,B}$ are positive. We first note that by the functional equation it is enough to verify positivity on the interval $[\frac{1}{\sqrt{N}}, \infty]$. Moreover, we have the estimate $|a_n| \leq n$ coming from the fact that each value of x can produce between 0 and 2 solutions for y. Taking a factor of q out of $g_{A,B}$ we see that $g_{A,B}$ is positive if and only if $1 + \sum_{k=2}^\infty a_k q^{k-1}$ is positive a condition which is guaranteed if $\sum_{k=2}^\infty kq^{k-1} = \frac{1}{(1-q)^2} - 1 \leq 1$, which leads to $q \leq 1 - \sqrt{\frac{1}{2}}$ or $t \geq \frac{-\log(1-\sqrt{\frac{1}{2}})}{2\pi} = 0.195....$ Consequently, it is enough to check positivity of $g_{A,B}/q$ on the interval $[\frac{1}{\sqrt{N}}, 0.196]$. This means that all curves with $N < 27$ automatically have positive $g_{A,B}$ functions. The first case which is not positive occurs when $N = 37$ for the simple reason that it has an infinite number of rational solutions. The first case of a non positive g when the number of solutions is finite occurs when $N = 203$. There are 1321 functions g with the corresponding equation having a finite number of solutions with conductor $N < 1000$. Together with A. Bond and S. Ryak, we checked positivity for all these functions and found that 1219 out of 1321 were positive. In the range $100,000 < N_E < 100,300$, we found 336 out of 441 to be positive, which is obviously a lower percentage but still a majority.

However, as in the case of the twisted theta functions it has been proved very recently, see [23], that only a sparse set of values A, B have positive $g_{A,B}$. In fact this is proved for many families of modular forms, including the Hecke newforms of a given weight with varying levels N. When we fix the level N and let $k \to \infty$, it has recently been shown, [19], that only a finite number of Hecke newforms g will be positive on the imaginary axis. See also [18] for a related result. For a more comprehensive account of these developments and the state of the art regarding the question whether there is an infinite number of Hecke newforms which are positive, see [35, 36].

References

[1] Bachmat E. and Sarfati H. 2010. Analysis of Size Interval Task Assignment `SITA` policies. *Performance Evaluation*, 67(2), 102-120, 2010.

[2] R. Baker and H. Montgomery, Oscillations of Quadratic L-functions, pages 23-40 *in Analytic Number Theory*, B. Brendt et. al. eds., Birkhauser, 1990.

[3] P. Bateman, G. Purdy and S. Wagstaff, Some numerical results on Fekete polynomials, *Math. Comp.*, collection of articles dedicated to Derrick Henry Lehmer on the occasion of his seventieth birthday, vol 29, 723, 1975.

[4] A. Beilinson, Higher regulators and values of L-functions. *Journal of Soviet Mathematics* 30 (1985), 2036-2070.

[5] A. Beilinson, Notes on absolute Hodge cohomology, in *Applications of algebraic K-theory to algebraic geometry and number theory*, Contemp. Math. 55, 3568, 1986.

[6] P. Biane, J. Pitman and M. Yor, Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions, Bull. Amer. Math. Soc. 38, 435-465, 2001.

[7] Birch,B ., Swinnerton-Dyer, H., Notes on elliptic curves II, *Journal reine u. angewandte Math.*. 218,79108, 1965.

[8] S. Bloch, Higher regulators, algebraic K-theory and zeta functions of elliptic curves, CRM monograph series vol. 11, AMS publications, 2000 (based on lectures given at Irvine U. in 1980.)

[9] S. Bloch, K. Kato, L-functions and Tamagawa numbers of motives, in *The Grothendieck Festschrift, Vol. I*, 333-400, Progr. Math., 86, Birkhauser Boston, Boston, MA, 1990.

[10] C. Breuil, B. Conrad, F. Diamond, and R. Taylor, On the modularity of elliptic curves over Q: wild 3-adic exercises, *Journal of the American Mathematical Society* 14 (2001), no. 4, 843-939.

[11] D. Bump, Automorphic forms and representations, Cambridge studies in advanced mathematics 55, Cambridge U. Press, 1996.

[12] S. Chowla and A. Selberg, On Epstein's zeta function. I, Proceedings of the National Academy of Sciences of the United States of America, 35, 371374, 1949.

[13] S. Chowla and A. Selberg, On Epstein's Zeta-function, Journal fur die reine und angewandte Mathematik 227, 86110, 1967.

[14] G. Cornell, J. Silverman and G. Stevens (Eds.), Modular forms and Fermat's last theorem, Springer Verlag, 1997.

[15] F. Diamond and J. Shurman, A first course in modular forms, Graduate texts in Math. 228, Springer verlag, 2005.

[16] P. Dirichlet, Beweis des Satzes, dass jede unbegrenzte arithmetische Progression, deren erstes Glied und Differenz ganze Zahlen ohne gemeinschaftlichen Factor sind, unendlich viele Primzahlen enthalt, *Abhand. Ak. Wiss. Berlin* 48, 1837.

[17] H. Feng, V. Misra and D. Rubenstein, Optimal state-free, size-aware dispatching for heterogeneous M/G/-type systems, *Performance evaluation*, vol. 62, 475-492, 2005.

[18] A. Ghosh, A. Reznikov, and P. Sarnak, On the nodal domains of maass forms, in preparation, 2012.

[19] A. Ghosh and P. Sarnak, Real zeros of holomorphic Hecke cusp forms, Arxiv:1103.3262, 2011.

[20] H. Hahn, On a conjecture of Fekete, *Journal of the Korean Math. Soc.*, vol 5, 13-16, 1968.

[21] M. Harchol-Balter, M. Crovella and C. Murta, On choosing a task assignment policy for a distributed server system, *IEEE Journal of parallel and distributed computing*, Vol. 59, 204-228, 1999.

[22] M. Harchol-Balter and R. Vesilo, To balance or unbalance load in size interval task allocation, *Probability in the Engineering and Informational Sciences*, to appear, 2010.

[23] J. Jung, On sparsity of positive definite automorphic forms within a family, in preparation, 2012.

[24] A. Khinchine, Mathematical theory of stationary queues, *Math. Sbornik*, 39, 73-84, 1932.

[25] J.F.C. Kingman, The single server queue in heavy traffic, *Proceedings of the Cambridge Philosophical Society*, vol. 57(4), 902-904, 1961.

[26] L. Kleinrock, *Queueing Systems. Volumes 1-2*, Wiley-Interscience, 1975.

[27] V. Kolyvagin, Finiteness of E(Q) and X(E,Q) for a class of Weil curves, *Math. USSR, Izv.*, 32, 523541, 1989.

[28] M. Mitzenmacher, Dynamic Models for File Sizes and Double Pareto Distributions, *Internet Mathematics*, vol 1(3), 305-334, 2004.

[29] M. Mitzenmacher A Brief History of Generative Models for Power Law and Lognormal Distributions *Internet Mathematics*, vol 1(2), 226-251, 2004.

[30] M. Mitzenmacher and B. Tworetzky, New Models and Methods for File Size Distributions, *In Proceedings of the 41st Annual Allerton Conference on Communication, Control, and Computing*, pp. 603612, Urbana, IL: University of Illinois at Urbana-Champaign, 2003.

[31] F. Pollaczek, Uber eine aufgabe dev wahrscheinlichkeitsthrorie I-II, *Math. Zeitschrift*, 32, 64-100 and 729-750, 1930.

[32] G. Polya, Verschiedene Bemerkungen zur Zahlentheorie, *Jahber. Deutsch. Math. Vereinigung*, vol 28, 31-40, 1919. Reprinted in Collected Papers, Vol III, MIT Press, Cambridge, Mass. 1984, pp. 76-85.

[33] W. J. Reed and M. Jorgensen. The Double Pareto- Lognormal Distribution - A New Parametric Model for Size Distributions. *Communications in Statistics: Theory and Methods* 33(8), 2004.

[34] W. Rudin, Real and complex analysis, McGraw-Hill, 1966.

[35] P. Sarnak, The distribution of mass and zeroes for high frequency eigenfunctions on the modular surface, *Slides from a talk given at the Dartmouth spectral geometry conference 2010*, available at http://www.math.dartmouth.edu/ specgeom/Sarnak.pdf

[36] P. Sarnak, positive definiteness of L-functions on the critical line, letter to E. Bachmat, available at www.cs.bgu.edu/ ebachmat

[37] C. L. Siegel, Lectures on advanced analytic number theory, Tata institute, 1961.

[38] J.H. Silverman, The arithmetic of elliptic curves, Graduate texts in mathematics 106, Springer Verlag, 1991.

[39] J.H. Silverman, Advanced topics in the arithmetic of elliptic curves, Graduate texts in mathematics 151, Springer Verlag, 1994.

[40] L. Takacs, A single server queue with Poisson input, *Operations research*, 10, 388-397, 1962.

[41] T. Tao, Benford's law, Zipf's law and the Pareto distribution, In "An epsilon of room II: pages from year three of a mathematical blog Terence Tao", AMS publications, 2011. Web version, terrytao.wordpress.com/2009/07/03.

[42] R. Taylor, A. Wiles, Ring-theoretic properties of certain Hecke algebras, *Annals of Math.*, 141, 553572, 1995.

[43] A. Weil, Uber die bestimmung Dirichletscher reihen durch funktionagleichungen, *Math. Annalen*, 168, 149-156, 1967.

[44] A. Wiles, Modular elliptic curves and Fermats last theorem, *Annals of Math.*, 142, 443551, 1995.