Integrable Systems and Combinatorial Theory

Percy Deift

The goal of this article is twofold:
- to describe the recent solution [BDJ1] of a problem in combinatorics using methods from the theory of integrable systems;
- to illustrate, en route, various developments that have taken place in the general area of integrable systems over the years.

The modern era in the theory of integrable systems began with the discovery in 1967 by Gardner, Greene, Kruskal, and Miura [GGKM] of a method for integrating the Korteweg de Vries equation. Initially this discovery was regarded as providing a method of solution restricted exclusively to the (rather thin) set of "integrable evolutionary systems", but by the early 1980s it started to become clear that the discovery of [GGKM] was just the first glimpse of a far more general *integrable method* that would eventually have applications across the broad spectrum of problems in pure and applied mathematics.

In the narrowest sense an *integrable system* is a Hamiltonian dynamical system—finite or infinite dimensional—with "enough" integrals of the motion, all of whose Poisson brackets are zero, to solve the system in some "explicit" form. It has been a broad and unifying experience in the field over the last thirty years that many systems that are of great mathematical and physical interest, which may not be Hamiltonian, and which may not even be dynamical, can be solved "explicitly" using techniques that have some direct link back to the method of solution for the Korteweg de Vries equation discovered in [GGKM]. This article, however, is not intended as a comprehensive bibliographical survey in the field. Descriptions of developments in integrable systems from the late 1960s to the early 1990s, together with the contributions of many different authors, can be found, for example, in [AS] and [FZ].

The problem at hand is the following. Let S_N be the group of permutations π of the numbers $1,\dots,N$. For $1 \le i_1 < \cdots < i_k \le N$ we say that $\pi(i_1),\dots,\pi(i_k)$ is an *increasing subsequence* of π of length k if $\pi(i_1) < \cdots < \pi(i_k)$. Let $\ell_N(\pi)$ be the length of the longest increasing subsequence of π. For example, if $N = 5$ and π is the permutation with $\pi(1) = 5$, $\pi(2) = 1$, $\pi(3) = 3$, $\pi(4) = 2$, and $\pi(5) = 4$, then 134 and 124 are both longest increasing subsequences of π and $\ell_5(\pi) = 3$. Equip S_N with the uniform distribution, and define

$$q_{n,N} \equiv \text{Prob}\,(\ell_N(\pi) \le n) = \frac{\#\text{ of }\pi\text{'s with }\ell_N(\pi) \le n}{N!}. \tag{1}$$

The basic problem is to determine the asymptotics of $q_{n,N}$ as $N \to \infty$. The recent survey [AD] is a useful reference for the properties of ℓ_N, various associated results, and some of the history to which we refer below.

Connections to Other Areas

Why is ℓ_N of interest? It turns out that there are connections to many different areas of mathematics.

Percy Deift is professor of mathematics at the University of Pennsylvania, and also at the Courant Institute of Mathematical Sciences, New York University. His e-mail addresses are deift@math.upenn.edu *and* deift@math1.cims.nyu.edu.

The author would like to thank Xin Zhou for many useful conversations and Craig Tracy for providing useful information. Special thanks are due to the editorial staff at the Notices *for all their wonderful and expert help in turning an unwieldy first draft into a viable text. This work was supported in part by NSF Grant DMS-9500867 and also by the Guggenheim Foundation.*

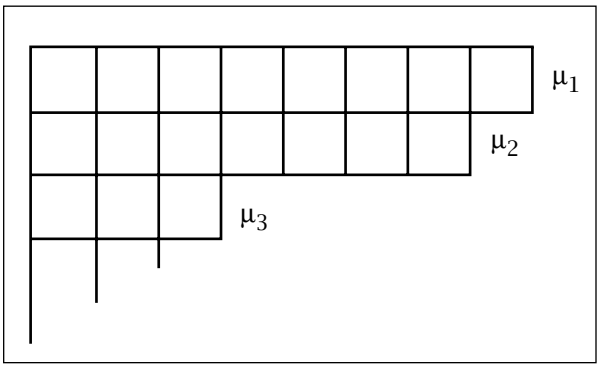

Figure 1. A Young diagram, or frame.

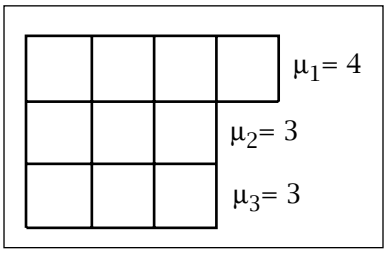

Figure 2. The Young diagram corresponding to $\mu = (4, 3, 3)$.

1	2	5	6
3	4	7	
8	9	10	

1	3	4	5
2	6	8	
7	9	10	

Figure 3. Two standard Young tableaux corresponding to the frame with $\mu = (4, 3, 3)$.

Representation Theory of S_N

The irreducible representations of S_N are parameterized by the partitions $\mu = (\mu_1, \ldots, \mu_p)$ of N,

$$\sum_{i=1}^{p} \mu_i = N\,, \qquad \mu_1 \geq \mu_2 \geq \cdots \geq \mu_p \geq 1.$$

Partitions $\mu = (\mu_1, \ldots, \mu_p)$ can in turn be represented by *Young diagrams*, or *frames*, as in Figure 1, where the first row contains μ_1 boxes, the second row contains μ_2 boxes, etc. For example, if $\mu = (4, 3, 3)$ is a partition of $N = 10$, then the associated frame is as in Figure 2. A standard Young tableau P is a one-to-one assignment of the numbers $1, 2, \ldots, N$ to the boxes in the given Young diagram or frame (denoted $Fr(P)$) in such a way that the numbers increase along the rows and down the columns. Figure 3 gives two examples of standard Young tableaux corresponding to the frame with $\mu = (4, 3, 3)$. The fact of the matter is that for any N, there is a bijection, the so-called *Robinson-Schented correspondence* RS, between S_N and pairs of standard Young tableaux with N boxes and with the same frame,

$$S_N \ni \pi \overset{\text{RS}}{\longmapsto} (P(\pi), Q(\pi)),$$
$$Fr(P(\pi)) = Fr(Q(\pi)).$$

Moreover, it turns out that

$$\ell_N(\pi) = \mu_1(P(\pi)) = \mu_1(Q(\pi)), \tag{2}$$

the number of boxes in the first row of $P(\pi)$ or $Q(\pi)$. Let f_μ denote the number of standard Young tableaux corresponding to a given frame μ with N boxes. Then by the RS-correspondence, the uniform distribution on S_N pushes forward to the so-called *Plancherel measure* on frames,

$$\text{Prob}(\{\mu\}) = \frac{f_\mu^2}{N!}\,.$$

It follows then by (2) that the analysis of $q_{n,N}$ is equivalent to a statistical problem in the "geometry" of the representations of S_N.[1]

Ulam's Metric

Consider a library of N books stacked, for simplicity, on one shelf. A deletion/insertion move consists of removing a book from the shelf and then returning it either by placing it at one of the ends of the stack or by inserting it between two books. Then a simple exercise shows that the quantity $N - \ell_N(\pi)$ is precisely the minimum number of deletion/insertion moves needed to place the library in linear order, $1, 2, \ldots, N$. This quantity in turn is the basic ingredient in the definition of *Ulam's metric*, $d(\pi, \sigma) = N - \ell_N(\pi\sigma^{-1})$ for π and σ in S_N. Ulam's metric is a bona fide metric on S_N and is commonly used in statistics to determine whether two permutations are "close".

Poisson Random Process

The statistics of ℓ_N are closely related to a certain Poisson process in the quadrant $\mathbb{R}_+ \times \mathbb{R}_+$. For this process, the number of particles N_A in a set $A \subset \mathbb{R}_+ \times \mathbb{R}_+$, has distribution given by $\text{Prob}(N_A = k) = e^{-|A|}\,|A|^k/k!$, $k = 0, 1, 2, \ldots$, where $|A|$ denotes the Lebesgue measure of A. Moreover, for any k, the particles $(x_1, y_1), \ldots (x_k, y_k)$ in A are independently and uniformly distributed. Let $x > 0$, and let L_x be the length of the longest upright path from $(0, 0)$ to (x, x), or more precisely, the largest integer k for which there are points $(x_i, y_i) \in \mathbb{R}_+ \times \mathbb{R}_+$ in the Poisson process with $0 < x_1 < \cdots < x_k < x$ and $0 < y_1 < \cdots < y_k < x$. This process, which is clearly closely related to ℓ_N, was in fact introduced by J. M. Hammersley in 1972 with the express purpose of studying ℓ_N. The advantage of the process is that the expected value $\mathbb{E}(L_x)$ can be analyzed for large x via subadditive ergodic theory, which then leads to information on $\mathbb{E}(\ell_N)$ as $N \to \infty$.

[1] *The reader who seeks more information about the irreducible representations of S_N may consult [Sa].*

Other applications of ℓ_N to statistical systems can be found, for example, in the recent work of Prähofer and Spohn (1999) on growth processes.

Patience Sorting

Patience sorting is a game played with a deck of cards labeled $1, 2, \ldots, N$. As described by Aldous and Diaconis in [AD], the game is played as follows.

> The deck is shuffled, cards are turned up one at a time and dealt into piles on the table, according to the following rules:
>
> A low card may be placed on a higher card (e.g. 2 may be placed on 7), or may be put into a new pile to the right of the existing piles.
>
> At each stage we see the top card on each pile. If the turned up card is higher than the cards showing, then it *must* be put into a new pile to the right of the others. The object of the game is to finish with as few piles as possible.

The *greedy* strategy is always to place a card on the leftmost possible pile. A simple computation shows that the greedy strategy is optimal. If the shuffled deck is in a "permutation state" π, we let $p_N(\pi)$ denote the number of piles one obtains by playing patience sorting starting from π and using the greedy strategy.

For example, suppose $N = 6$ cards are in the order

$$4 \quad 1 \quad 3 \quad 5 \quad 6 \quad 2$$

(corresponding to the permutation $\pi(1) = 4$, $\pi(2) = 1$, ...). Then the game proceeds as follows:

	1	1	1	1	1 2
4	4	4 3	4 3 5	4 3 5 6	4 3 5 6

and $p_6(\pi) = 4$. An elementary argument establishes the main fact of the matter, namely $p_N(\pi) = \ell_N(\pi)$. Indeed, in the above example we see that 1356 is the longest increasing subsequence for π, and so $\ell_N(\pi) = 4 = p_6(\pi)$.

Patience sorting, however, is more than just a game. Indeed, as $p_N(\pi) = \ell_N(\pi)$, patience sorting, in fact, provides a viable algorithm to compute ℓ_N. Also, in another direction, it is possible to interpret the p_N-problem, and hence the ℓ_N-problem, as a continuous space analog of so-called "first passage oriented percolation" (see [AD]).

Random Topologies on Surfaces

In a remarkable recent development Okounkov [Ok] has shown that the statistics of ℓ_N is also related to the problem of random topologies on surfaces (see below).

Previous Results

What is known about ℓ_N? As in so many combinatorial problems, the path leads back to Erdős. In 1935 Erdős and Szekeres proved the following Ramsey-theory-type result: every sequence of N distinct real numbers contains an increasing subsequence and/or a decreasing subsequence of length at least $\sqrt{N}$. As the increasing and decreasing subsequences are clearly distributed in the same way, it follows that the expected value $\mathbb{E}(\ell_N)$ satisfies

$$\mathbb{E}(\ell_N) \geq \tfrac{1}{2}\sqrt{N}.$$

After his experiences during the war with computing and computing machines, Ulam was constantly on the lookout for good computational problems to try out on the latest machines as they were developed. Ulam and Erdős had a long and enduring friendship, and one can only conjecture that at some point in their discussions the question arose about the computation of $\mathbb{E}(\ell_N)$. In 1961, using Monte Carlo techniques, Ulam computed $\mathbb{E}(\ell_N)$ for N lying in the (then very large!) range $1 \leq N \leq 10$ and found that

$$\mathbb{E}(\ell_N) \sim 1.70\sqrt{N}. \tag{3}$$

With characteristic boldness, Ulam viewed (3) as an asymptotic result and conjectured that

$$\lim_{N\to\infty} \frac{\mathbb{E}(\ell_N)}{\sqrt{N}} = c \tag{4}$$

exists. What became known as Ulam's Problem was to prove that the limit in (4) indeed exists and to compute the constant c. Subsequent numerical work of Baer and Brock in 1968 suggested that the value of c was 2. The first analytical result was due to Hammersley, who established in 1972 that the limit in (4) indeed exists, but he obtained only limited information on c. As noted in the previous section, Hammersley studied $\mathbb{E}(\ell_N)$ by introducing an associated Poisson process: the existence of the limit (4) then follows from subadditive ergodic theory. Finally, in 1977, Logan and Shepp showed that $c \geq 2$, and, independently, Vershik and Kerov proved that $c = 2$, thus solving Ulam's Problem. Alternative solutions have now been given by Aldous-Diaconis, T. Sepäläinen, and K. Johansson.

The question then arose about other statistics for ℓ_N, such as the variance $\mathrm{Var}(\ell_N)$. It was believed that as $N \to \infty$, $\mathrm{Var}(\ell_N) \sim N^{\alpha}$ for some α, and there were various conjectures for the value of α, including the value $\alpha = 1/3$ by H. Kesten, based on arguments from the theory of first passage percolation. In the early 1990s A. Odlyzko decided to use large-scale computations to clarify the situation. Together with E. Rains, Odlyzko performed Monte Carlo simulations using values of $N \leq 10^9$. These simulations indicated that

$$\lim_{N\to\infty} \frac{1}{N^{1/3}} \mathrm{Var}(\ell_N) = c_0$$

for some numerical constant $c_0 \sim 0.819$, which agrees, in particular, with the value $\alpha = 1/3$. Odlyzko and Rains also computed $\mathbb{E}(\ell_N)$ to higher order and found

$$\lim_{N\to\infty} \frac{\mathbb{E}(\ell_N) - 2\sqrt{N}}{N^{1/6}} = c_1 \ ,$$

where $c_1 \sim -1.758$. By the mid to late 1990s, apart from various large deviation results, this was basically where the matter stood.

Main Results

Before stating the results in [BDJ1] we need some notation and definitions. Let $Ai(x)$ denote the classical Airy function, given by

$$(Ai)_{xx} = xAi, \quad Ai(x) \sim \frac{e^{-\frac{2}{3}x^{3/2}}}{2\sqrt{\pi}\, x^{1/4}} \text{ as } x \to +\infty.$$

Then let $u(x)$ be the solution of the Painlevé II equation:

$$u_{xx} = 2u^3 + xu \tag{5}$$

with $u(x) \sim -Ai(x)$ as $x \to +\infty$. As shown by Hastings and McLeod (1980), this solution exists for all x and is unique, and the asymptotics as $x \to \pm\infty$ are given by

(6)
$$\begin{cases} u(x) = -Ai(x) + O\Big(\dfrac{e^{-\frac{4}{3}x^{3/2}}}{x^{1/4}}\Big) & \text{as } x \to +\infty \\ \quad = -\sqrt{\dfrac{-x}{2}}\,\Big(1 + O\Big(\dfrac{1}{x^2}\Big)\Big) & \text{as } x \to -\infty. \end{cases}$$

We will say more about the history and general properties of Painlevé functions further on. Define the Tracy-Widom distribution, introduced in [TW], by

$$F(t) = e^{-\int_t^\infty (x-t)u(x)^2\,dx}. \tag{7}$$

Observe from (6) and (7) that $F'(t) > 0$, $F(t) \to 1$ as $t \to +\infty$, and $F(t) \to 0$ as $t \to -\infty$; thus F is indeed a distribution function. Let χ be a random variable with distribution F, so that $\text{Prob}(\chi \le t) = F(t)$. Define χ_N by centering and scaling ℓ_N as follows:

$$\chi_N(\pi) = \Big(\ell_N(\pi) - 2\sqrt{N}\Big)\Big/ N^{1/6}.$$

Then we have the following result.

Theorem 1 *[BDJ1]. Let S_N be the group of permutations of N numbers with uniform distribution, and let $\ell_N(\pi)$ be the length of the longest increasing subsequence of $\pi \in S_N$. Then as $N \to \infty$,*

$$\chi_N \to \chi \qquad \text{in distribution}$$

i.e.

$$\lim_{N\to\infty} \text{Prob}\,(\chi_N \le t) = F(t).$$

Theorem 2 gives convergence of the moments.

Theorem 2 *[BDJ1]. For any $m = 0, 1, 2, \ldots$,*

$$\lim_{N\to\infty} \mathbb{E}(\chi_N^m) = \mathbb{E}(\chi^m). \tag{8}$$

Consequently

$$\lim_{N\to\infty} \frac{\text{Var}(\ell_N)}{N^{1/3}} = \int t^2\,dF(t) - \Big(\int t\,dF(t)\Big)^2 \tag{9}$$

and

$$\lim_{N\to\infty} \frac{\mathbb{E}(\ell_N) - 2\sqrt{N}}{N^{1/6}} = \int t\,dF(t). \tag{10}$$

If one solves the Painlevé II equation for $u(x)$ numerically and then computes the integrals on the right-hand side of (9) and (10) approximately, one obtains the values 0.8132 and -1.7711, which are in good agreement with the Monte Carlo calculations of Odlyzko and Rains for c_0 and c_1 given above.

Random Matrix Theory

It turns out that there is a very interesting connection between the above results and the Gaussian Unitary Ensemble (GUE) of random matrix theory.[2] GUE denotes the ensemble of $N \times N$ Hermitian matrices $M = (M_{ij})$ with probability distribution

$$z_N^{-1} e^{-\text{Tr}\,M^2}\,dM = z_N^{-1} e^{-\text{Tr}\,M^2} \prod_{i=1}^N dM_{ii} \times \prod_{1\le i<j\le N} d(\text{Re}\,M_{ij})\,d(\text{Im}\,M_{ij}),$$

where z_N is a normalization coefficient. Random matrix theory was introduced into theoretical physics by E. Wigner in the 1950s in his work on the scattering resonances of neutrons off large nuclei. With the work of Dyson, Gaudin, Mehta, Wigner and many others, random matrix theory has since developed into a major tool of theoretical physics: one way to measure the extent of the activity in the subject is to look at the recent review of C. W. J. Beenakker (1997) containing over 400 papers, mostly from the 1980s and 1990s, devoted, almost exclusively, just to the applications of random matrix theory to the problem of quantum transport. What was not clear, however, at least at the beginning, was that random matrix theory also had a role to play in pure mathematics. But all this changed with a remarkable computation of H. E. Montgomery. In the early 1970s, assuming the Riemann Hypothesis, Montgomery began a series of calculations on the statistics of the zeros $\left\{\frac{1}{2} + i\gamma_j\right\}$ of the Riemann zeta function, leading, in particular, to a formula R for the two-point correlation function for the (appropriately rescaled) γ_j's. What happened next is one of the most celebrated denouements in mathematics in recent years: In a well-known and oft-repeated

[2] *A basic reference for random matrices is [Me].*

conversation, F. Dyson, upon hearing about Montgomery's work, immediately wrote down a formula for the two-point function and asked Montgomery, "Did you obtain this?" Montgomery was astounded and wanted to know how Dyson knew the answer. "Well," said Dyson, "if the zeros of zeta behave like the eigenvalues of a random matrix, then the two-point function would have to be R!" Montgomery's formula R was indeed precisely the two-point function for the eigenvalues of a random GUE matrix! Suddenly, the scattering of neutrons off nuclei had something to do with the zeros of the Riemann zeta function. There have been many beautiful and extraordinary extensions of Montgomery's work, as can be seen from the recent monograph (1999) of N. Katz and P. Sarnak.

It turns out that the result of Montgomery concerns the behavior of the eigenvalues of a random matrix in the bulk of the spectrum. What concerns us here, however, is the behavior of the eigenvalues at an edge of the spectrum. In 1994 Tracy and Widom [TW] considered the largest eigenvalue $\lambda_1(M)$ of an $N \times N$ GUE matrix M. After centering and scaling they showed that

$$\lim_{N\to\infty} \text{Prob}\Big(\Big(\lambda_1(M) - \sqrt{2N}\Big)\sqrt{2}\,N^{\frac{1}{6}} \le t\Big) = F(t),$$

where $F(t)$ is precisely the distribution given in (7). In other words, the result in [BDJ1] shows that the length ℓ_N of the longest increasing subsequence of a random permutation behaves statistically like the largest eigenvalue of a random GUE matrix as $N \to \infty$.

In [TW] the authors also computed the limiting statistics of the k^{th} eigenvalue $\lambda_k(M)$ of a random GUE matrix, $\lambda_1(M) \ge \lambda_2(M) \ge \cdots \ge \lambda_k(M) \ge \ldots$:

$$\lim_{N\to\infty} \text{Prob}\Big(\Big(\lambda_k(M) - \sqrt{2N}\Big)\sqrt{2}\,N^{\frac{1}{6}} \le t\Big) = F_k(t),$$
$$k \ge 1,$$

and they expressed each $F_k(t)$ in terms of solutions of the Painlevé II equation (in this notation $F(t) = F_1(t)$). The question then arose in [BDJ1] whether any object in the permutation picture behaved statistically like $\lambda_2(M)$, say, the second largest eigenvalue. A moment's reflection shows that it cannot be the length of the second longest increasing subsequence: if we drop one element in a longest increasing subsequence, we obtain a second longest increasing subsequence, which would then be distributed in the same way as ℓ_N. In order to arrive at a reasonable guess for λ_2, recall the Robinson-Schensted correspondence mentioned above. By (2)

$$\ell_N(\pi) = \mu_1(P(\pi)) = \mu_1(Q(\pi)),$$

the number of boxes in the first row of $P(\pi)$ or $Q(\pi)$. Motivated by this result, the conjecture in [BDJ1] became that $\mu_2(P(\pi)) = \mu_2(Q(\pi))$, the number of boxes in the second row of $P(\pi)$ or $Q(\pi)$, behaves statistically like $\lambda_2(M)$ as $N \to \infty$, or, more generally,

$$\mu_k(P(\pi)) \quad \text{behaves statistically like} \quad \lambda_k(M)$$

for each $k \ge 1$ as $N \to \infty$. The analog of (8) above for m equal to 1 and 2 was then verified numerically in [BDJ1] for λ_2 against computations of Odlyzko and Rains. The subsequent paper [BDJ2] contains the proof of the analogs of Theorems 1 and 2 above for λ_2. Over the year 1999 three different groups—[Ok], [BOO], and [Jo]—have independently proved the above conjecture for the (joint distribution of the) first k eigenvalues for arbitrary $k \ge 1$. The proof in [BOO] is representation theoretic, while the proof in [Jo] involves ideas from statistical mechanics together with the asymptotics of polynomials orthogonal with respect to a certain distinguished discrete measure. The proofs in [BDJ1], [BDJ2], [BOO], and [Jo] are calculations: one computes the asymptotics of the distribution for the μ_k's and then compares them with the Tracy-Widom asymptotics for the λ_k's. The proof in [Ok], however, provides a picture relating random matrices and random permutations. The context for the connection is the problem of random topologies on surfaces (see above). On the one hand, one can view a surface as glued together from polygons by identifying sides of polygons in pairs. Such representations are called *maps* on the surface. Via quantum gravity, it has long been known that maps are intimately connected with random matrix theory. On the other hand, one can realize a surface as a ramified covering of the sphere S^2. It is a classical fact, however, that any problem about the combinatorics of coverings is equivalent to a problem about permutations. In other words, there are two different ways to produce a random surface: as $N \to \infty$ these two ways become equivalent, and this leads to the relation between the λ_k's and the μ_k's.

Beginning in late 1998 there has been a surge of activity in this area, and many papers by many different authors have appeared relating combinatorics and random matrix theory. We refer the reader to [BR] for a very recent survey.

Steps in the Proofs of Theorems 1 and 2

The remainder of this article is devoted to describing the main ideas in the proofs of Theorems 1 and 2 in [BDJ1]. As we will see, the key analytic fact is that the problem can be rephrased as a *Riemann-Hilbert Problem* (RHP) with oscillatory/exponential factors of the kind that can be analyzed by the noncommutative *steepest descent method* introduced by Xin Zhou and the author in [DZ].

Step 1: Poissonization

For $\lambda \in \mathbb{R}$, set

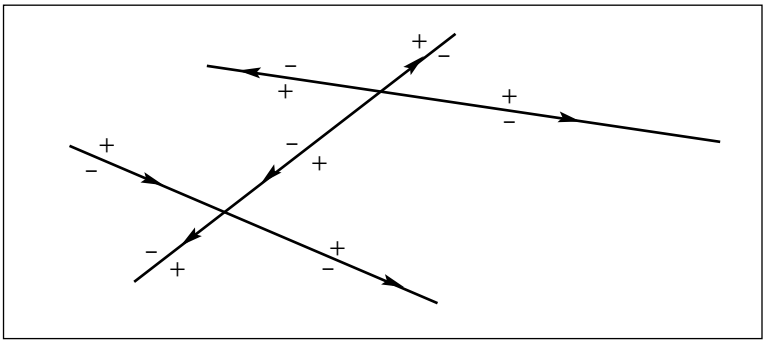

Figure 4. Oriented contour Σ.

$$\phi_n(\lambda) = \sum_{N=0}^{\infty} \frac{e^{-\lambda}\lambda^N}{N!}\, q_{n,N}$$

where $q_{n,N}$ is the distribution function for ℓ_N given in (1) above. It turns out that $q_{n,N}$ is monotonically decreasing in N, and Johansson (1998) showed that this fact leads to a strong Tauberian-like "de-Poissonization" result. In turn this result reduces the problem of the asymptotics for $q_{n,N}$ to computing the asymptotics of $\phi_n(\lambda)$ for $n \le N$ with $\lambda \sim N \to \infty$. A priori we have just substituted one asymptotic problem for another, but Poissonization helps because of the following wonderful fact: there is an exact formula for $\phi_n(\lambda)$, namely

$$\phi_n(\lambda) = e^{-\lambda} D_{n-1}(\lambda), \tag{11}$$

where $D_{n-1}(\lambda)$ is the $n \times n$ Toeplitz determinant (see, e.g., [Sz]) generated by the function $e^{2\sqrt{\lambda}\cos\theta}$, i.e.,

$$D_{n-1}(\lambda) = \det\left(c_{jk} = c_{k-j} = \int_{-\pi}^{\pi} e^{-i(k-j)\theta} e^{2\sqrt{\lambda}\cos\theta}\, \frac{d\theta}{2\pi}\right)_{0\le k,j\le n-1}.$$

Formula (11) was derived by I. Gessel in 1990, and new proofs have now been given by many different authors (see [BDJ1] for references). The asymptotics of Toeplitz determinants D_n as $n \to \infty$ have been studied intensively over the years, starting with the famous Szegö strong-limit theorem in 1952; the novel feature here is that n and λ tend to ∞, a so-called double-scaling limit.

Step 2: Connection to Orthogonal Polynomials

Let $p_n(z) = \kappa_n(\lambda) z^n + \ldots, \kappa_n(\lambda) > 0$ be the n^{th} orthonormal polynomial with respect to the weight $e^{2\sqrt{\lambda}\cos\theta}\, \frac{d\theta}{2\pi}$ on the circle, i.e.,

$$\int_{-\pi}^{\pi} \overline{p_n(e^{i\theta})}\, p_m(e^{i\theta}) e^{2\sqrt{\lambda}\cos\theta}\, \frac{d\theta}{2\pi} = \delta_{n,m}, \qquad n, m \ge 0.$$

Then standard calculations as in [Sz] show that $\kappa_n^2(\lambda) = D_{n-1}(\lambda)/D_n(\lambda)$, which then leads to the formula

$$\log \phi_n(\lambda) = \sum_{k=n}^{\infty} \log \kappa_k^2(\lambda).$$

Step 3: Riemann-Hilbert Problem

It is at this point that we make contact with one of the central techniques of integrable systems, namely the Riemann-Hilbert problem (RHP).[3] The RHP and related techniques (see [FZ]) were introduced into integrable theory by A. B. Shabat in 1976 and were developed intensively from the early 1980s onwards, particularly by the group around A. Its in Leningrad. A crucial step in the development of RHP as a powerful tool in the rigorous analysis of inverse scattering problems was the seminal work of Beals and Coifman (1984) on first-order systems of ordinary differential equations on the line. Over the years an extraordinary variety of problems in pure and applied mathematics have been expressed in terms of an RHP. Typically these RHPs contain oscillatory or exponentially growing/decreasing factors such as e^{izx} or e^{-iz^2t}, z lying on a curve in $\mathbb{C}$, and the asymptotic behavior of the solution of the problem as x and t tend to ∞ can be computed by using, for example, the steepest descent method introduced in [DZ] and extended as described in Step 5 below. Following Fokas, Its, and Kitaev (1991), we show here, among other things, how to compute $\kappa_k^2(\lambda)$ in terms of the solution of an RHP.

What is an RHP? Suppose we have an oriented contour Σ in $\mathbb{C}$ as in Figure 4. By convention, if we traverse an arc in Σ in the direction of the arrow, we say that the + (resp. −) side lies to the left (resp. right). Suppose in addition that we have a map $v : \Sigma \to GL(k, \mathbb{C})$ with $v, v^{-1} \in L^\infty(\Sigma)$. Then the RHP (Σ, v) consists of the following: find an $\ell \times k$ matrix-valued function $m = m(z)$ such that

- m is analytic in $\mathbb{C} \setminus \Sigma$;
- $m_+(z) = m_-(z)v(z)$ for almost every $z \in \Sigma$, where $m_\pm(z) \equiv \lim_{\substack{z' \to z \\ z' \in (\pm) \text{ side}}} m(z')$.

If in addition $\ell = k$ and

- $m(z) \to I$ as $z \to \infty$,

we say that the RHP is normalized at ∞. The study of the RHP (Σ, v) can be reduced to the analysis of a singular integral equation on Σ. The question of the existence and uniqueness of the solution of the RHP, the precise sense in which the boundary values $m_\pm$ are attained on Σ, and the rate at which $m(z) \to I$ as $z \to \infty$ in the normalized case are subtle issues that have been studied extensively over the years, and we refer the reader to the literature for details.

We now introduce the RHP for orthogonal polynomials discovered by Fokas, Its, and Kitaev. Let Σ be the unit circle in $\mathbb{C}$, oriented counterclockwise, as in Figure 5. For k and λ fixed let

[3] *A standard reference for RHPs is [CG].*

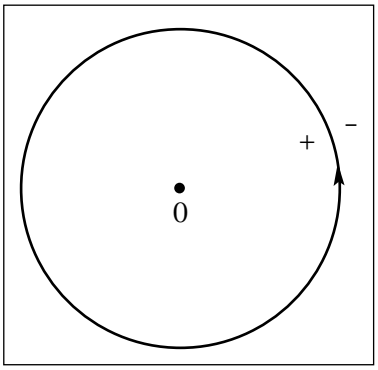

Figure 5. $\Sigma = \{z : |z| = 1\}$.

$Y(z;k+1,\lambda) = (Y_{ij}(z;k+1,\lambda))_{1\le i,j\le 2}$ be the 2×2 matrix function solving the following RHP:

(12)
$$\begin{cases} \bullet\ Y(z;k+1,\lambda) \quad \text{is analytic in } \mathbb{C}\setminus\Sigma \\ \bullet\ Y_+(z;k+1,\lambda) = Y_-(z;k+1,\lambda)\begin{pmatrix} 1 & z^{-(k+1)}e^{\sqrt{\lambda}(z+z^{-1})} \\ 0 & 1 \end{pmatrix} \\ \quad \text{for } z\in\Sigma, \\ \bullet\ Y(z;k+1,\lambda)\begin{pmatrix} z^{-(k+1)} & 0 \\ 0 & z^{k+1} \end{pmatrix} \to I \quad \text{as } z\to\infty. \end{cases}$$

Then

$$\kappa_k^2(\lambda) = -Y_{21}(z=0;k+1,\lambda).$$

Also $p_{k+1}(z) = \kappa_k^2(\lambda)Y_{11}(z;k+1,\lambda)$.

So to evaluate $\kappa_k^2(\lambda)$ for $k \ge n$, and hence $\phi_n(\lambda)$ and then $q_{n,N}$, we must control the solution Y of the above RHP in the limit when the two parameters k and $\sqrt{\lambda}$ in the multiplier $z^{-(k+1)}e^{\sqrt{\lambda}(z+z^{-1})}$ are large. As indicated earlier, this is precisely the kind of situation that can be controlled by the noncommutative steepest descent method.

Step 4: Painlevé Theory

Here we make contact with another central tool in the theory of integrable systems, namely nonlinear special function theory, and in particular the Painlevé equations.[4] It is increasingly recognized that the Painlevé equations play a role in modern mathematical physics analogous to the role of the classical special functions in the nineteenth century. For example, the Painlevé equations appear as self-similar solutions of the Korteweg de Vries equation, as correlation functions for the transverse Ising chain in the infinite temperature limit, and also in nonperturbative two-dimensional quantum gravity, amongst many other physical applications.

The origin of the Painlevé equations is rather esoteric. Towards the end of the nineteenth century, Painlevé considered the following problem: Let $y = y(z;a,b)$ be the solution of the equation

(13)
$$\frac{d^2y}{dz^2} = F\left(z, y, \frac{dy}{dz}\right)$$

[4] *The classical theory of the Painlevé equations is described in the book of E. L. Ince,* Ordinary Differential Equations, *and a comprehensive survey of relatively recent results and applications can be found in [FZ].*

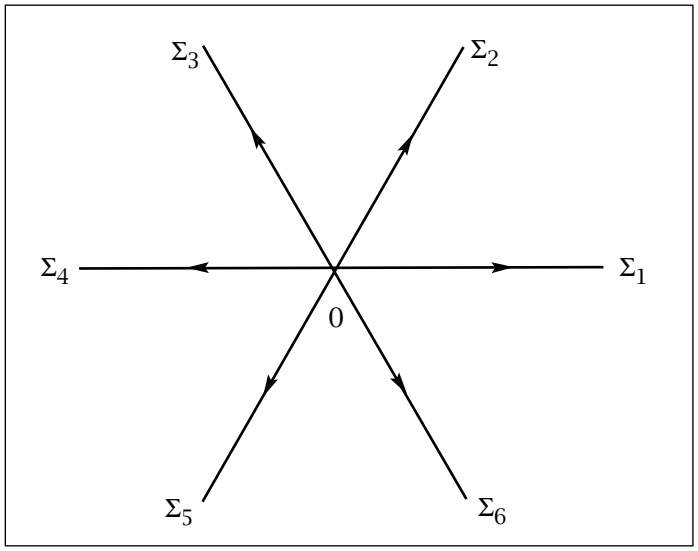

Figure 6. Σ **for the Painlevé II equation.**

with initial data

(14)
$$y\Big|_{z_0} = a, \qquad \frac{dy}{dz}\Big|_{z_0} = b.$$

Here

(15)
$$\begin{gathered} F \text{ is meromorphic in } z \\ \text{and rational in } y \text{ and } dy/dz. \end{gathered}$$

Painlevé asked the following question: for which functions F satisfying (15) do all solutions $y = y(z;a,b)$ of (13) and (14) have the following property? As the initial data a,b vary, the only singularities of $y = y(z;a,b)$ that can move are poles. In other words, any branch points or essential singularities of $y(z;a,b)$ must remain fixed as a and b vary. Colloquially, one says that solutions of (13) have only "movable poles". Quite remarkably, Painlevé and Gambier found that, modulo changes of dependent and independent variables, there were precisely fifty such equations (13). Forty-four of these could be reduced to known equations, such as the hypergeometric equation, etc., but the remaining six could not. They became known as the six Painlevé transcendents. Apart from efforts by a few isolated mathematicians, work on Painlevé theory ceased after the First World War. A number of the young researchers working in the area were killed in battle, and Painlevé himself went on to a distinguished political career in France. Then in the mid-1960s, after lying essentially dormant for almost fifty years, Painlevé functions suddenly started to appear as solutions of a variety of problems in engineering and mathematical physics. At this stage it is quite remarkable, and a little mysterious, how many problems in pure and applied mathematics involve Painlevé functions in their solution. From the arcane origins of the subject, it would have been hard to predict that the Painlevé equations would have had such a universal impact.

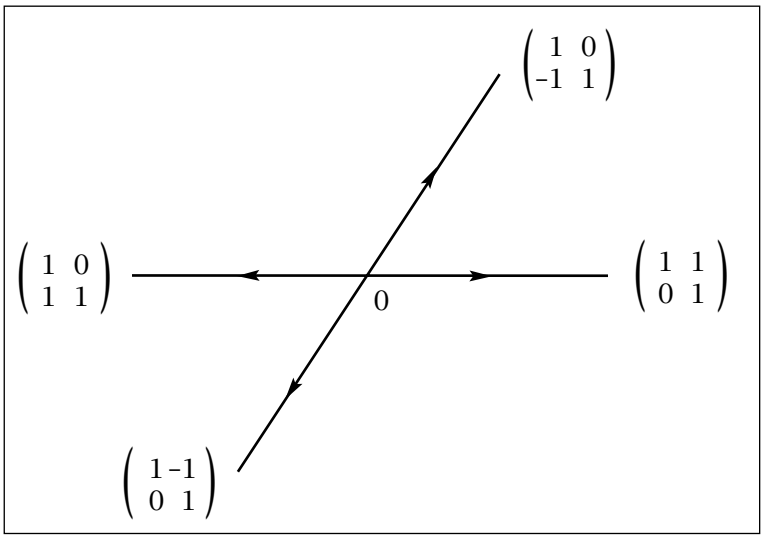

Figure 7. Σ for the Painlevé II equation with $p = -q = 1$ and $r = 0$.

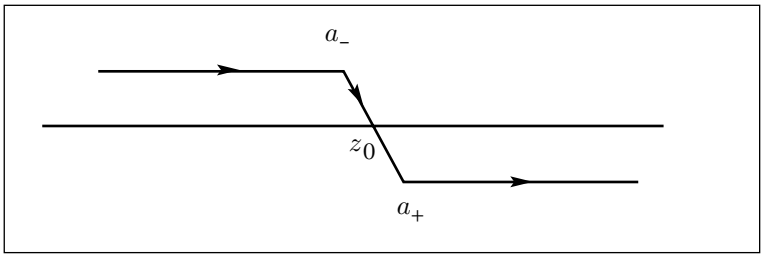

Figure 8. Deformed contour C_ϵ, $a_\pm = z_0 \pm \epsilon(1 - i)$.

The modern theory of Painlevé equations rests on the discovery of Flaschka-Newell (1980) and Jimbo-Miwa-Ueno (1980) independently that the Painlevé equations can be solved in terms of an associated RHP. We describe the situation for the Painlevé II equation (5), $u_{xx} = 2u^3 + xu$, which is of special interest in [BDJ1] and [BDJ2]. Here the contour is a union of six rays,

$$\Sigma = \bigcup_{k=1}^{6} \left\{ \Sigma_k = e^{i(k-1)\pi/3}\, \mathbb{R}_+ \right\},$$

as in Figure 6. The underlying jump matrix $v = v(z)$ for the problem is constant on each of the rays $\Sigma_1, \Sigma_2, \ldots, \Sigma_6$ and is given by

$$\begin{pmatrix} 1 & p \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ q & 1 \end{pmatrix}, \begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ p & 1 \end{pmatrix}, \begin{pmatrix} 1 & q \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ r & 1 \end{pmatrix}$$

respectively, where the constants p, q, r are related through

$$p + q + r + pqr = 0. \tag{16}$$

For $x \in \mathbb{R}$, set

$$v_x(z) = e^{-i(4z^3/3+xz)\sigma_3} v(z) e^{i(4z^3/3+xz)\sigma_3}, \qquad z \in \Sigma,$$

where σ_3 is the Pauli matrix $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. The result of Flaschka-Newell and Jimbo-Miwa-Ueno is the following: Let $m = m(z; x)$ be the 2×2 matrix solution of the normalized RHP (Σ, v_x). If we write

$$m(z; x) = I + \frac{m_1(x)}{z} + O\left(\frac{1}{z^2}\right) \qquad \text{as } z \to \infty,$$

then

$$u(x) \equiv 2i(m_1(x))_{12}$$

is a solution of the Painlevé II equation (5). By (16) the RHP has two free parameters, and so we obtain a general solution of (5). The particular solution of (5) that occurs in the Tracy-Widom distribution (7) corresponds to the choice $p = -q = 1$, $r = 0$. In this case Σ_3 and Σ_6 may be removed from Σ, and we are led to an RHP of the form in Figure 7.

Step 5: Steepest Descent Method

Building on the work of many authors from the 1970s and 1980s, in particular the work of Zakharov-Manakov, Segur-Ablowitz, and Its-Novokshenov, new techniques were introduced in the 1990s that made it possible to analyze in great detail the asymptotic behavior of integrable systems as external parameters in the system, say space and time, become large. An analogy is in order: The Schrödinger equation

$$iq_t + q_{xx} = 0, \qquad q(x, 0) = q_0(x),$$

can be solved via the transform

$$q(x, t) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{i(xz - tz^2)} \hat{q}_0(z)\, dz \tag{17}$$

where $\hat{q}_0$ is the Fourier transform of the initial data q_0. But in order to determine the fine structure of the solution $q(x, t)$, one must analyze in detail the mapping properties of the above transform (or more properly of the *Fourier integral operator* with multiplier $e^{i(xz-tz^2)}$) as $|x|$ and t tend to ∞. The main contribution to (17) comes from the *point of stationary phase* of $e^{i(xz-tz^2)}$, i.e., the point $z = z_0$ for which $\frac{d}{dz}(xz - tz^2) = 0$. Clearly $z_0 = x/2t$. One way to see the importance of z_0—the so-called *method of stationary phase*—is to note that for z away from z_0,

$$e^{i(xz-tz^2)} = e^{itz_0^2} e^{-it(z-z_0)^2}$$

oscillates rapidly in z for $t \gg 1$. Consequently, at least formally, the contribution of $\hat{q}_0(z)$ to the integral cancels out as $t \to \infty$, except for an arbitrary small neighborhood $\{|z - z_0| < \epsilon\}$ of $z = z_0$. Alternatively, if $\hat{q}_0(z)$ is analytic in a strip, say $\{|\mathrm{Im}\, z| < 2\epsilon\}$, we can use Cauchy's Theorem to deform (17) to an integral on a contour C_ϵ as in Figure 8. A simple calculation now shows that $\left| e^{i(xz-tz^2)} \right| = \left| e^{-it(z-z_0)^2} \right|$ decreases rapidly on C_ϵ away from $z = z_0$, and as $t \to \infty$ we see again that the problem localizes to a neighborhood of $z = z_0$. In this second approach—the so-called *method of steepest descent*—the analysis is considerably simplified: we do not need to keep track

of delicate cancellations, and absolute estimates suffice.

By contrast with the linear case, the integrable method applied to the nonlinear Schrödinger equation

$$iq_t + q_{xx} - 2|q|^2 q = 0, \qquad q(x,0) = q_0(x),$$

leads to a formula for the solution $q(x,t)$ of the form

$$q(x,t) = S^{-1}\left(e^{ix(\cdot)-t(\cdot)^2} r(\cdot)\right) \tag{18}$$

where $r(z)$ corresponds to the initial data q_0 and S^{-1}, the so-called *inverse scattering map*, is a highly nonlinear functional of its argument. Formula (18) was first obtained by Zakharov-Shabat in 1971. When r (or equivalently q_0) is small in an appropriate sense, (18) reduces to the Fourier integral operator (17). This then raises the question whether there exists an analog of the above stationary phase/steepest descent methods that can be applied directly to S^{-1} in order to determine the asymptotic properties of $q(x,t)$ as $|x|$ and t tend to ∞. Most importantly, the method should not require r (or equivalently q_0) to be small. The steepest descent-type method introduced in [DZ] is precisely such a scheme. The method was developed further by the authors and their collaborators through the 1990s. In particular, together with S. Venakides (1997), the method was extended in an essential way in order to treat a broad class of problems with fully nonlinear oscillations. The method in [DZ], as well as its various extensions, is based on a reformulation of the problem at hand as an RHP and proceeds by deforming contours much in the spirit of the classical method of steepest descent as described above. The fundamental difference is that here we have a matrix problem and the method has noncommutative and nonlinear elements that lead to new phenomena. We refer the reader to [DKMVZ1], [DKMVZ2], and the references therein for a detailed description of the method and its development.

The details of how the steepest descent method is applied to the RHP (12) in order to prove Theorems 1 and 2 above are closely related to [DKMVZ1] and [DKMVZ2] and are unfortunately beyond the scope of this article. We will, however, give some indication how the Painlevé II equation enters the problem. The critical region for the RHP is where $2\sqrt{\lambda}/(k+1)$ is comparable with 1, which splits into the two parts, where $2\sqrt{\lambda}/(k+1) \le 1$ and where $2\sqrt{\lambda}/(k+1) \ge 1$. We write these sets as $2\sqrt{\lambda}/(k+1) \lesssim 1$ and $2\sqrt{\lambda}/(k+1) \gtrsim 1$. For $2\sqrt{\lambda}/(k+1) \lesssim 1$, the RHP is deformed to an equivalent RHP (Σ', v') on a contour Σ' as in Figure 9, the original contour $\Sigma = \{|z| = 1\}$ being indicated by dashed lines. As λ and k tend to ∞ with $2\sqrt{\lambda}/(k+1) \lesssim 1$, the jump matrix v' converges uniformly and exponentially

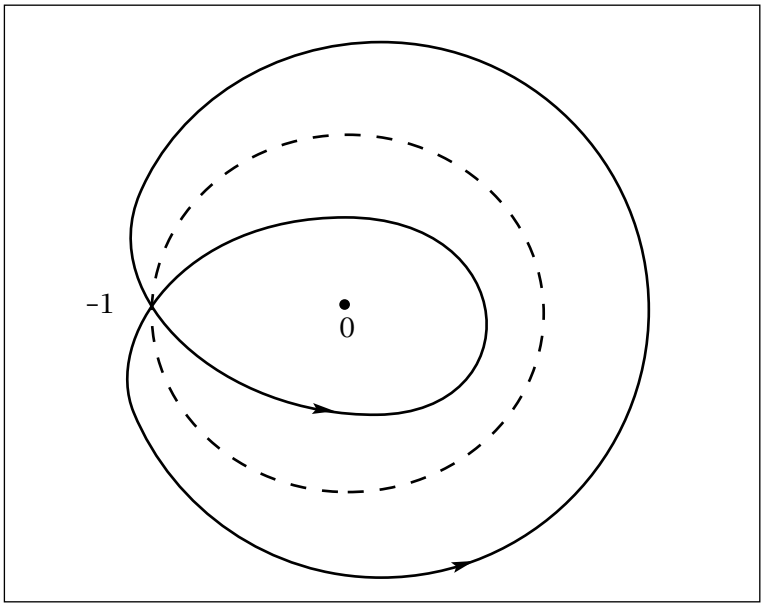

Figure 9. Deformed contour Σ'.

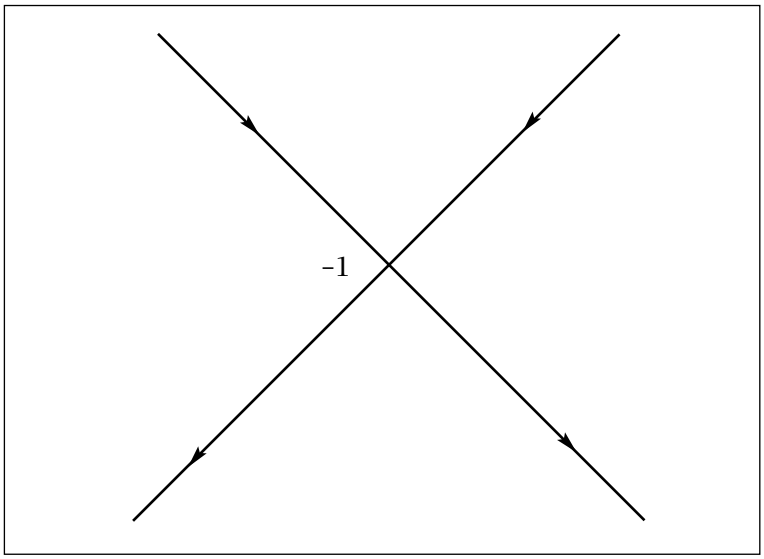

Figure 10. Limiting contour.

to the identity matrix on compact subsets of $\Sigma' \setminus \{-1\}$. This means that the RHP (Σ', v') reduces to an RHP on a small cross around $z = -1$, as in Figure 10. After scaling and rotation, we obtain precisely the RHP shown in Figure 7, and this is how the Painlevé II equation enters the problem. There is a similar deformation for the case that λ and k tend to ∞ with $2\sqrt{\lambda}/(k+1) \gtrsim 1$.

As noted above, in [BDJ2] the authors prove the analogs of Theorems 1 and 2 for the second eigenvalue λ_2. The key new problem in [BDJ2] is to find an analog of Gessel's formula (11) for the Poissonization $\widetilde{\phi}_n(\lambda)$ of the distribution function $\widetilde{q}_{n,N}$ for the number of boxes in the second row of a standard Young tableau with Plancherel measure,

$$\widetilde{\phi}_n(\lambda) = \sum_{N=0}^{\infty} \frac{e^{-\lambda}\lambda^N}{N!}\, \widetilde{q}_{n,N}.$$

Here we make contact with yet another central technique in integrable theory, the concept of an *integrable operator*. Let Σ be an oriented contour in $\mathbb{C}$. We say that an operator K acting on $L^2(\Sigma)$ is *integrable* if it has a kernel of the form

$$K(z,z') = \frac{\sum_{j=1}^{k} f_j(z) g_j(z')}{z - z'}, \qquad z, z' \in \Sigma,$$

for some functions f_j and g_j with $1 \le j \le k < \infty$. Special examples of integrable operators began to appear in field theory, in statistical mechanical models, and also in random matrix theory in the 1960s and 1970s, particularly in the work of Wu-McCoy-Tracy-Baruch and later Sato-Miwa-Jimbo. Integrable operators as a distinguished class of operators of independent interest, however, were first singled out by Its-Izergin-Korepin-Slavnov in 1990; for a recent review, see Deift (1999).

Integrable operators form an algebra, but the most striking property of these operators is that the inverse $(1-K)^{-1}$ can be expressed explicitly in terms of the solution of a canonically associated RHP on Σ. It turns out that the solution of a wide variety of problems in mathematics and mathematical physics can be expressed as a determinant $\det(1-K)$ for some integrable operator K. As

$$\log\det(1-K) = -\int_0^1 \operatorname{Tr}\left(\frac{1}{1-tK}\, K\right) dt,$$

it follows that the solution of these problems can be expressed explicitly in terms of an associated RHP. Typically, K depends on certain external physical parameters, and the asymptotic behavior of the system as these parameters become large can then be inferred by applying the steepest descent method to the RHP. This is true in particular for $\widetilde{\phi}_n(\lambda)$ above as λ and n tend to ∞: here Σ is the unit circle in $\mathbb{C}$ and the underlying integrable operator has the form

$$K(z,z') = \frac{z^{-n}(z')^n - \varphi(z)(\varphi(z'))^{-1}}{z - z'},$$

where $\varphi(z) = e^{\sqrt{\lambda}(z-z^{-1})}$. The details are given in [BDJ2].

References

[AD] D. Aldous and P. Diaconis, Longest increasing subsequences: From patience sorting to the Baik–Deift–Johansson theorem, *Bull. Amer. Math. Soc. (N.S.)* **36** (1999), 413–432.

[AS] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM Studies in Applied Mathematics, vol. 4, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.

[BDJ1] J. Baik, P. Deift, and K. Johansson, On the distribution of the length of the longest increasing subsequence of random permutations, *J. Amer. Math. Soc.* **12** (1999), 1119–1178.

[BDJ2] ____, On the distribution of the length of the second row of a Young diagram under Plancherel measure, LANL e-print Archive `math.CO/9901118; http://xxx.lanl.gov/`; to appear in *Geom. and Funct. Anal.*

[BOO] A. Borodin, A. Okounkov, and G. Olshanski, On asymptotics of Plancherel measures for symmetric groups, LANL e-print Archive `math.CO/990532; http://xxx.lanl.gov/`.

[BR] J. Baik and E. M. Rains, Symmetrized random permutations, LANL e-print Archive `math.CO/9910019; http://xxx.lanl.gov/`.

[CG] K. Clancey and I. Gohberg, *Factorization of Matrix Functions and Singular Integral Operators*, Operator Theory: Advances and Applications, vol. 3, Birkhäuser-Verlag, Basel and Boston, 1981.

[DKMVZ1] P. Deift, T. Kriecherbauer, K. T-R. McLaughlin, S. Venakides, and X. Zhou, Strong asymptotics of orthogonal polynomials with respect to exponential weights, *Comm. Pure Appl. Math.* **52** (1999), 1491–1552.

[DKMVZ2] ____, Uniform asymptotics for polynomials orthogonal with respect to varying exponential weights and applications to universality questions in random matrix theory, *Comm. Pure Appl. Math.* **52** (1999), 1335–1425.

[DZ] P. Deift and X. Zhou, A steepest descent method for oscillatory Riemann-Hilbert problems. Asymptotics for the MKdV equation, *Ann. of Math.* **137** (1993), 295–368.

[FZ] A. S. Fokas and V. E. Zakharov (eds.), *Important Developments in Soliton Theory*, Springer-Verlag, Berlin and Heidelberg, 1993.

[GGKM] C. S. Gardner, J. M. Greene, M. D. Kruskal, and R. M. Miura, Method for solving the Korteweg-de Vries equation, *Phys. Rev. Lett.* **19** (1967), 1095–1097.

[Jo] K. Johansson, Discrete orthogonal polynomial ensembles and the Plancherel measure, preprint (1999).

[Me] M. L. Mehta, *Random Matrices*, 2nd edition, Academic Press, Boston, 1991.

[Ok] A. Okounkov, Random matrices and random permutations, LANL e-print Archive `math.CO/9903176; http://xxx.lanl.gov/`.

[Sa] B. Sagan, *The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1991.

[Sz] G. Szegö, *Orthogonal Polynomials*, 4th edition, Colloquium Publications, vol. 23, Amer. Math. Soc., Providence, RI, 1975.

[TW] C. A. Tracy and H. Widom, Level-spacing distributions and the Airy kernel, *Comm. Math. Phys.* **159** (1994), 151–174.