Rationality and meromorphy of zeta functions

Alan G.B. Lauder *

January 7, 2005

1 Introduction

This article is all about two theorems on equations over finite fields which have been proved in the past decade. Here they are:

Theorem 1.1. *The rigid cohomology of a variety over a finite field is finite dimensional.*

Theorem 1.2. *The unit root zeta function of a family of varieties over a finite field of characteristic p is p-adic meromorphic.*

The purpose of the article is to explain what these theorems mean, and also to give an outline of the proof of the first one. The intended audience is mathematicians with an interest in finite fields, but no especial expertise on the vast literature which surrounds the topic of equations over finite fields. By way of motivation, we will begin by giving an indication of the historical significance of these two theorems, before giving more formal definitions in Section 2.

The basic object of interest to us is a system of polynomial equations over a finite field. Loosely speaking, this is called a *variety*. Given such a system, one can encode the number of solutions over different finite extension of the base field in a generating function. This is the *zeta function* of the variety. In the late 1950s Dwork proved that this generating function is always a rational function. Weil had conjectured this some ten years earlier, and conceived a plan for proving it based upon an as yet unknown *cohomology theory* for varieties over finite fields. Such a theory would associate a vector space with a variety over a finite field, and the rationality of zeta functions would follow from the finite dimensionality of these vector spaces. To everyone's surprise, Dwork proved rationality without constructing such a theory. He proved instead that the zeta function was *meromorphic* as a p-adic function, and then deduced that it must be rational. During the next decade Dwork's work inspired the construction of a true cohomology theory based upon p-adic analysis. Unfortunately though,

*Mathematical Institute, Oxford University, 24-29 St Giles, Oxford OX1 3LB. E-mail: lauder@maths.ox.ac.uk. The author is a Royal Society University Research Fellow. He wishes to thank Kiran Kedlaya, Jonathan Pila and Daqing Wan for helpful comments on the article.

no one could prove the vector spaces it associated to varieties were finite dimensional. Theorem 1.1 solves this problem, and thus gives the first p-adic cohomological proof of the rationality of zeta functions. Dwork's work in the 1960s also led him to associate zeta functions with families of varieties over finite fields. The most important were the *unit root* zeta functions. Dwork conjectured that these mysterious functions were p-adic meromorphic — they are known, though, not to be rational in general. His own techniques were inadequate for proving this conjecture, and the cohomological machinery being developed at the time was geared up to proving functions were rational. Theorem 1.2 settles Dwork's conjecture.

The paper is organised in the following manner. We will begin in Section 2 by outlining the meaning of Theorem 1.1. Finer details and an explanation of the proof will be given via the study of an explicit surface. This surface is introduced in Section 3. Sections 4 and 5 more-or-less give a proof of the finiteness of the cohomology of our surface. This serves as a model for the proof in the general case, which is sketched in Section 6. In Section 7 we use our surface to explain Theorem 1.2. We will not give much idea as to how it is proved, but we will be able to explain why one cannot prove it using the methods Dwork used to show the rationality of zeta functions. We conclude in Section 8 by attributing the various results and techniques in this paper. For now we shall just mention that special cases and generalisations of Theorem 1.1 have been proved by Berthelot [2], Grosse-Klönne [7], Kedlaya [10], Mebkhout [15] and Tsuzuki [20]. Theorem 1.2 is due to Wan [23, 24, 25].

2 Zeta functions and cohomology

We begin by noting that Theorems 1.1 and 1.2 are true for arbitrary varieties over finite fields; however, we shall restrict our attention solely to the case of affine varieties, as these are simpler to work with, and this turns out to be the essential case anyway.

Let $\mathbb{F}_q$ be the finite field with q elements of characteristic p, and denote by $\bar{\mathbb{F}}_q$ an algebraic closure of $\mathbb{F}_q$. Let $\bar{\mathcal{X}}$ be an affine variety over $\mathbb{F}_q$. Thus $\bar{\mathcal{X}}$ is defined by the common vanishing of a collection of polynomials $f_1, \ldots, f_m \in \mathbb{F}_q[X_1, \ldots, X_n]$ for some m and n. The ring $\bar{A} := \mathbb{F}_q[X_1, \ldots, X_n]/(f_1, \ldots, f_m)$ is called the *coordinate ring* of $\bar{\mathcal{X}}$. Formally $\bar{\mathcal{X}} := \mathrm{Spec}(\bar{A})$. For each integer $k \geq 1$, let $\mathbb{F}_{q^k} \subset \bar{\mathbb{F}}_q$ be the unique subfield of order q^k. The set of $\mathbb{F}_{q^k}$-rational points on $\bar{\mathcal{X}}$ is denoted $\bar{\mathcal{X}}(\mathbb{F}_{q^k})$ and has cardinality $|\bar{\mathcal{X}}(\mathbb{F}_{q^k})|$. Thus $\bar{\mathcal{X}}(\mathbb{F}_{q^k})$ is the set of points $(x_1, \ldots, x_n) \in \mathbb{F}_{q^k}^n$ where all of the polynomials $f_1, \ldots, f_m$ vanish.

We can now define the main object of interest.

Definition 2.1. *The zeta function of $\bar{\mathcal{X}}$ is the formal power series*

$$Z(\bar{\mathcal{X}}, T) := \exp\left(\sum_{k \geq 1} |\bar{\mathcal{X}}(\mathbb{F}_{q^k})| \frac{T^k}{k}\right) \in \mathbb{Q}[[T]].$$

Weil conjectured and Dwork proved that this is a rational function [4]. Specifically

$$Z(\bar{\mathcal{X}}, T) = \frac{P(T)}{Q(T)},\ P, Q \in 1 + T\mathbb{Z}[T],\ \gcd(P, Q) = 1.$$

So factoring the numerator and denominator one has

$$\exp\left(\sum_{k\geq 1} |\bar{\mathcal{X}}(\mathbb{F}_{q^k})| \frac{T^k}{k}\right) = \frac{\prod_i (1 - \alpha_i T)}{\prod_j (1 - \beta_j T)}.$$

Taking the logarithmic derivatives of both sides and equating powers of T one finds that

$$|\bar{\mathcal{X}}(\mathbb{F}_{q^k})| = \sum_j \beta_j^k - \sum_i \alpha_i^k \text{ for } k \geq 1,$$

which is an attractive and useful formula.

In the 1960s inspired by Dwork's work, Monsky and Washnitzer constructed a functor which associates with each *smooth* affine variety $\bar{\mathcal{X}}$ over $\mathbb{F}_q$ a vector space $H^*_{MW}(\bar{\mathcal{X}})$. (Smoothness just means that the matrix of partial derivatives $(\frac{\partial f_j}{\partial X_i})$ has full rank when evaluated at any point on the variety.) The vector space is defined over the field $\mathbb{Q}_q$; this is the unramified extension of degree $\log_p(q)$ of the field of p-adic numbers $\mathbb{Q}_p$. See [13] for details on these fields. The essential point is that $\mathbb{Q}_q$ has characteristic zero, and so contains a copy of $\mathbb{Q}$. The vector space decomposes as $H^*_{MW}(\bar{\mathcal{X}}) := \oplus_{i=0}^{\dim(\bar{\mathcal{X}})} H^i_{MW}(\bar{\mathcal{X}})$. Here $\dim(\bar{\mathcal{X}})$ is the dimension of $\bar{\mathcal{X}}$; assuming $f_1, \ldots, f_m$ are sufficiently generic this is just $n - m$. On each of these vector spaces there is a linear operator Frob_q called the *Frobenius*. (We shall see explicit examples of these spaces and this operator in Section 5.) Monsky proved a formula [18]

$$Z(\bar{\mathcal{X}}, T) = \prod_{i=0}^{\dim(\mathcal{X})} \det(1 - Tq^{\dim(\bar{\mathcal{X}})} \mathrm{Frob}_q^{-1} | H^i_{MW}(\bar{\mathcal{X}}))^{(-1)^{i+1}}. \tag{2.1}$$

Assuming the spaces $H^i_{MW}(\bar{\mathcal{X}})$ are finite dimensional, this gives a cohomological proof of the rationality of $Z(\bar{\mathcal{X}}, T)$. Unfortunately it was not known whether these spaces were finite dimensional. (The formula does not assume this, as Monsky was able to make sense of the determinants for infinite dimensional spaces.) We shall sketch a proof of the following theorem.

Theorem 2.2. *The spaces $H^i_{MW}(\bar{\mathcal{X}})$ are finite dimensional.*

The functor $\bar{\mathcal{X}} \mapsto H_{MW}(\bar{\mathcal{X}})$ is called *Monsky-Washnitzer cohomology*. It is only defined for smooth affine varieties; however, nowadays it is a special case of a more general theory due to Berthelot called *rigid cohomology* which is defined for arbitrary varieties. So Theorem 2.2 is a special case of Theorem 1.1. We shall focus on Theorem 2.2 for the rest of the paper.

3 A surface fibred into smooth curves

We now introduce the example which will be used throughout the article to explain the meaning of both Theorems 1.1 and 1.2 and give an idea of the proof of the first. It is a surface in affine 3-space which has a very convenient fibration into smooth curves. In the next three sections, we shall explain how the rigid cohomology of this surface is defined, and how its finiteness can be proved via the fibration. This technique illustrates the key induction step in the proof of finiteness for general smooth affine varieties. The induction argument we present in Section 6 actually takes us outside of the category of smooth affine varieties, and into a larger category of *overconvergent F-isocrystals* defined on such varieties. Our sketch-proof will actually show that *their* cohomology is finite dimensional.

Here is our example: Let $\bar{\mathcal{X}} = \mathrm{Spec}(\bar{A})$ where

$$\bar{A} := \mathbb{F}_q[X, Y, \Gamma, Y^{-1}, \bar{r}(\Gamma)^{-1}]/(Y^2 - \bar{Q}(X, \Gamma)).$$

Here $\bar{Q}(X, \Gamma) \in \mathbb{F}_q[X, \Gamma]$ is monic in X of degree $2g + 1$, and q is a power of an odd prime p. We define

$$\bar{r}(\Gamma) := \mathrm{Res}(\bar{Q}, \frac{\partial \bar{Q}}{\partial X}, X) \in \mathbb{F}_q[\Gamma],$$

the Sylvester resultant with respect to X of the polynomials $\bar{Q}, \frac{\partial \bar{Q}}{\partial X} \in \mathbb{F}_q[\Gamma][X]$ [3, Pages 150-151]. This is the determinant of a matrix over $\mathbb{F}_q[\Gamma]$ formed by extracting the coefficients of powers of X in the two polynomials. The polynomial $\bar{r}(\Gamma)$ vanishes precisely at the elements $\bar{\gamma} \in \bar{\mathbb{F}}_q$ for which $Q(X, \bar{\gamma})$ is not squarefree. We *assume* that the polynomial $\bar{r}(\Gamma)$ is not identically zero. We shall write $\sqrt{\bar{Q}}$ for Y, and so

$$\bar{A} = \left\{ \sum_m \sum_{i=0}^{2g} \frac{a_{m,i}(\Gamma) X^i}{\sqrt{\bar{Q}}^m} \mid a_{m.i} \in \mathbb{Q}_q[\Gamma, \bar{r}(\Gamma)^{-1}] \right\}. \tag{3.1}$$

Here the sum over $m \in \mathbb{Z}$ is finite. Set-theoretically, $\bar{\mathcal{X}}$ is just the set of solutions in the affine 3-space to the polynomial system: $Y^2 = \bar{Q}(X, \Gamma)$, $Y \neq 0, \bar{r}(\Gamma) \neq 0$.

For each $\bar{\gamma} \in \bar{\mathbb{F}}_q$, let $\bar{\mathcal{X}}_{\bar{\gamma}}$ denote the curve over $\mathbb{F}_q(\bar{\gamma})$ defined as $\mathrm{Spec}(\bar{A}_{\bar{\gamma}})$ where

$$\bar{A}_{\bar{\gamma}} := \mathbb{F}_q(\bar{\gamma})[X, Y, Y^{-1}]/(Y^2 - \bar{Q}(X, \bar{\gamma})).$$

Set-theoretically, this curve is just the points on the affine hyperelliptic curve $Y^2 = \bar{Q}(X, \bar{\gamma})$ with $Y \neq 0$. The affine hyperelliptic curve $Y^2 = \bar{Q}(X, \bar{\gamma})$ is smooth precisely when $\bar{Q}(X, \bar{\gamma})$ is squarefree. Therefore, the affine hyperelliptic curve is smooth if and only if $\bar{r}(\bar{\gamma}) \neq 0$. The curve $\bar{\mathcal{X}}_{\bar{\gamma}}$ is this hyperelliptic curve with the *ramification points* removed; thus the map $(x, y) \mapsto x$ makes it an unramified cover of the affine line with the roots of $\bar{Q}(X, \bar{\gamma})$ removed. Although we will not refer to it explicitly, it is this nice map, coupled with the smoothness of the affine hyperelliptic curve $Y^2 = \bar{Q}(X, \bar{\gamma})$, which makes the construction of the cohomology spaces for $\bar{\mathcal{X}}_{\bar{\gamma}}$ particularly simple when $\bar{r}(\bar{\gamma}) \neq 0$.

Let $\bar{\mathcal{S}} := \mathrm{Spec}(\bar{B})$ where $\bar{B} := \mathbb{F}_q[\Gamma, \bar{r}(\Gamma)^{-1}]$. So $\bar{\mathcal{S}}$ is the affine line with the roots of $\bar{r}(\Gamma)$ removed. Applying the "Spec" functor to the embedding $\bar{B} \to \bar{A}$ gives the family

$$\bar{f} : \bar{\mathcal{X}} \to \bar{S}$$

This is the fibration of our surface into curves. Since we have removed the roots of $\bar{r}(\Gamma)$ from the base of this fibration, all of the fibres are smooth and remain smooth when their ramification points are replaced. Formally, the fibres are $\bar{\mathcal{X}}_{\bar{\gamma}} = \bar{\mathcal{X}} \times \mathbb{F}_q(\bar{\gamma})$ where the fibre product is via the specialisation map $\Gamma \mapsto \bar{\gamma}$.

We will construct the Monsky-Washnitzer (a.k.a. rigid) cohomology $H^*_{MW}(\bar{\mathcal{X}})$ of the surface $\bar{\mathcal{X}}$, and we shall use the fibration to show that $H^2_{MW}(\bar{\mathcal{X}})$ is finite dimensional.

4 de Rham cohomology of a lifting

We first introduce some notation for p-adic numbers, see [13]. Recall $\mathbb{Q}_q$ is the unramified extension of $\mathbb{Q}_p$ of degree $\log_p(q)$. Let $\mathbb{Z}_q$ be the ring of integers of $\mathbb{Q}_q$. There is a reduction modulo p map $\mathbb{Z}_q \to \mathbb{F}_q$. Let $\mathbb{C}_p$ denote a completion of an algebraic closure of $\mathbb{Q}_p$. Let $\mathrm{ord} : \mathbb{C}_p \to \mathbb{Q}$ be the p-adic order map, normalised so that $\mathrm{ord}(p) = 1$. Write $\mathbb{O}_p$ for the ring of integers of $\mathbb{C}_p$, i.e., elements of non-negative p-adic order.

4.1 The cohomology of our surface

Our first step in the construction is to lift the surface $\bar{\mathcal{X}}$ to characteristic zero. This is quite simple: Define

$$A := \mathbb{Q}_q[X, Y, \Gamma, Y^{-1}, r(\Gamma)^{-1}]/(Y^2 - Q(X, \Gamma)).$$

Here $Q(X, \Gamma) \in \mathbb{Z}_q[X, \Gamma]$ is any polynomial which is monic in X of degree $2g+1$ and reduces to $\bar{Q}$ modulo p. We have

$$r(\Gamma) := \mathrm{Res}(Q, \frac{\partial Q}{\partial X}, X) \in \mathbb{Z}_q[\Gamma],$$

which reduces to $\bar{r}(\Gamma)$ modulo p. Elements in A are exactly as in (3.1), only with $\bar{Q}$, $\bar{r}$ and $\mathbb{F}_q$ replaced by Q, r and $\mathbb{Q}_q$. Let $\mathcal{X}$ be the subset of points $(x, y, \gamma) \in \mathbb{O}_p^3$ which reduce modulo p to points on $\bar{\mathcal{X}}$. Notice that $\mathcal{X}$ is independent of our choice of Q.

Now that we are in characteristic zero, there is a construction called *algebraic de Rham cohomology* which associates in a functorial manner a finite dimensional vector space $H^*_{dR}(\mathcal{X})$ with A. First, one first constructs the module $\Omega(A/\mathbb{Q}_q)$ of derivations of A over $\mathbb{Q}_q$. A *derivation* of A over $\mathbb{Q}_q$ is a $\mathbb{Q}_q$-linear map d from A to an A-module which satisfies the Leibniz rule $d(ab) = adb + bda$ for all $a, b \in A$. The module $\Omega(A/\mathbb{Q}_q)$ comes equipped with such a derivation $d : A \to \Omega(A/\mathbb{Q}_q)$, and it is universal in the sense that any other derivation must factor through it. In our case, looking at (3.1) it is apparent there are

only two "independent derivations": differentiation by X and Γ. So $\Omega(A/\mathbb{Q}_q)$ is the free A-module generator by "symbols" dX and $d\Gamma$, and

$$d : g \mapsto \frac{\partial g}{\partial X} dX + \frac{\partial g}{\partial \Gamma} d\Gamma \text{ for } g \in A.$$

The second step is to construct the *de Rham complex* from the exterior powers of $\Omega(A/\mathbb{Q}_q)$. In our case this complex is

$$0 \to A \xrightarrow{d_0} AdX + Ad\Gamma \xrightarrow{d_1} AdXd\Gamma \to 0$$

Here $d_0 = d$ and

$$d_1 : g_1 dX + g_2 d\Gamma \mapsto \left(\frac{\partial g_1}{d\Gamma} - \frac{\partial g_2}{\partial X} \right).$$

The de Rham cohomology spaces are

$$H^0_{dR}(\mathcal{X}) := \ker(d_0),\ H^1_{dR} := \ker(d_1)/\mathrm{im}(d_0),\ H^2_{dR}(\mathcal{X}) := AdXd\Gamma/\mathrm{im}(d_1).$$

We hope that these $\mathbb{Q}_q$-vector spaces are finite dimensional. This is certainly the case for $H^0_{dR}(\mathcal{X})$ since the only functions which map to zero are the constants $\mathbb{Q}_q$. We shall pass over $H^1_{dR}(\mathcal{X})$ and focus on the top space $H^2_{dR}(\mathcal{X})$. Our aim is to understand why this is finite dimensional.

The space $\mathrm{im}(d_1)$ is the set 2-forms $rdXd\Gamma$ which are the sum of a 2-form which can be "formally integrated" with respect to X and one which can be "formally integrated" with respect to Γ. Thus the quotient represents 2-forms which cannot be broken up in this way and formally integrated. We would like to find a finite set of 2-forms such that every 2-form can be written as a linear combination of these, plus one which can be broken into two pieces and each piece formally integrated. Thinking about both X and Γ at the same time is a little difficult. Let's consider integration by X first of all.

4.2 The relative cohomology of the family

Define $B = \mathbb{Q}_q[\Gamma, r(\Gamma)^{-1}]$ and let $\mathcal{S}$ be the subset of points $\gamma \in \mathbb{O}_p$ with $r(\gamma) \neq 0 \bmod p$. This is lifting of the base of our fibration. We have a family $f : \mathcal{X} \to \mathcal{S}$ in characteristic zero. Forgetting about Γ amounts formally to considering the *relative de Rham cohomology* of this family of curves. We shall write this as $H^*_{dR}(\mathcal{X}/\mathcal{S})$. This is constructed as before, only this time we forget about Γ and consider B-linear derivations, i.e., derivations which kill Γ. The *module of relative differentials* $\Omega(A/B)$ is an A-module which encodes all of these, and comes equipped with a universal derivation $\partial : A \to \Omega(A/B)$. Since we only have differentiation with respect to X left, we find $\Omega(A/B) = AdX$ with $\partial : g \to \frac{\partial g}{\partial X} dX$. The de Rham complex is now just

$$0 \to A \xrightarrow{\partial} AdX \to 0.$$

We are interested in the quotient $H^1_{dR}(\mathcal{X}/\mathcal{S}) := AdX/\mathrm{im}(\partial)$. This is much easier to work with, and we can see quite easily it is a finitely generated module over B.

Specifically, we claim that $H^1_{dR}(\mathcal{X}/\mathcal{S})$ is spanned as a module over B by the forms

$$\left\{ \frac{X^i dX}{\sqrt{Q^j}} \,|\, j = 1 \text{ and } 0 \le i < 2g,\ j = 2 \text{ and } 0 \le i \le 2g \right\}. \tag{4.1}$$

One can reduce elements of AdX to linear combinations of the forms (4.1) modulo $\mathrm{im}(\partial)$ as follows. Write $Q' = \frac{\partial Q}{\partial X}$. For $P \in \mathbb{Q}_q[X, \Gamma]$ using the Sylvester matrix [3, Pages 150-151] and some linear algebra we can write $r(\Gamma)P = R_0 Q + S_0 Q'$ for some polynomials $R_0, S_0 \in \mathbb{Z}_q[X, \Gamma]$ whose degrees may be explicitly bounded. So $P = RQ + SQ'$ where $R = R_0/r(\Gamma)$ and $S = S_0/r(\Gamma)$ have coefficients in B. For $m \ge 1$

$$\partial\left(\frac{S}{Q^{m/2}}\right) = \frac{S'dX}{Q^{m/2}} - \frac{mSQ'dX}{2Q^{m/2+1}}.$$

Hence in homology:

$$\begin{aligned} \frac{PdX}{Q^{m/2+1}} &= \frac{(RQ + SQ')dX}{Q^{m/2+1}} \\ &\equiv \frac{RdX}{Q^{m/2}} + \frac{2S'dX}{mQ^{m/2}}. \end{aligned} \tag{4.2}$$

Iterating this relation an appropriate number of times can reduce any form to the shape $*dX/Q^{j/2}$, for $j = 1, 2$ and $* \in \mathbb{Q}_q[\Gamma, r(\Gamma)^{-1}][X]$. Reduction of $*$ to a polynomial of the appropriate degree in X is easier: A form $*dX/\sqrt{Q}$ with $*$ of degree $m \ge 2g$ can be reduced in degree by subtracting an appropriate "B-multiple" of $\partial(X^{m-2g}\sqrt{Q})$; a form $*dX/Q$ with $*$ of degree $m > 2g$ can be reduced in degree by subtracting an appropriate multiple of $\partial(X^{m-2g})$. So this shows that forms can be reduced to B-linear combinations of our spanning set (4.1), and so certainly the quotient $H^1_{dR}(\mathcal{X}/\mathcal{S})$ is finitely generated. In fact, the quotient $H_{dR}(\mathcal{X}/\mathcal{S})$ is a free B-module of rank $4g + 1$, although we shall not prove this.

4.3 Application of a "spectral sequence"

To see how this is related to $H^2_{dR}(\mathcal{X})$ consider the commutative square.

$$\begin{array}{ccccccc} 0 & \longrightarrow & A & \xrightarrow{\frac{\partial \cdot}{\partial X} dX} & AdX & \longrightarrow & 0 \\ & & \downarrow \frac{\partial \cdot}{\partial \Gamma} d\Gamma & & \downarrow \frac{\partial \cdot}{\partial \Gamma} d\Gamma & & \\ 0 & \longrightarrow & Ad\Gamma & \xrightarrow{\frac{\partial \cdot}{\partial X} dX} & AdXd\Gamma & \longrightarrow & 0. \end{array} \tag{4.3}$$

Writing $H^1_{dR}(\mathcal{X}/\mathcal{S})d\Gamma$ for the cokernel of the bottom map, by commutativity we have that $\frac{\partial \cdot}{\partial \Gamma} d\Gamma$ induces a map

$$\nabla : H^1_{dR}(\mathcal{X}/\mathcal{S}) \to H^1_{dR}(\mathcal{X}/\mathcal{S})d\Gamma.$$

Explicitly, given a B-linear combination of the spanning forms (4.1) the map ∇ differentiates it with respect to Γ and then reduces it back to a B-linear

combination of the spanning forms. The map ∇ is called a *connection*. It is additive and satisfies the Leibniz rule

$$\nabla(bm) = \frac{\partial b}{\partial \Gamma} m d\Gamma + b\nabla(m)$$

for any $b \in B$ and $m \in H^1_{dR}(\mathcal{X}/\mathcal{S})$.

Elements in $H^1_{dR}(\mathcal{X}/\mathcal{S})d\Gamma$ represent 2-forms which have been "reduced with respect to X", i.e., and appropriate 2-form which is the derivative with respect to X of a 1-form $*d\Gamma$ has been subtracted to put it in a nice form. One would now like to reduce these 2-forms with respect to Γ. Specifically, consider the quotient

$$\operatorname{coker}(\nabla) = H^1_{dR}(\mathcal{X}/\mathcal{S})d\Gamma/\nabla(H^1_{dR}(\mathcal{X}/\mathcal{S})).$$

Showing that this space is a finite dimensional $\mathbb{Q}_q$-vector space will imply $H^2_{dR}(\mathcal{X})$ is also finite dimensional. More precisely, the two spaces are isomorphic. Formally, this isomorphism arises from a *spectral sequence* associated to the fibration.

4.4 Finiteness of H^1 of a D-module

The space $\operatorname{coker}(\nabla)$ is an example of the first homology space of a *D-module*. A technique for proving this is finite dimensional was given by Monsky [19, Lemma 5] based on an idea of J.C. Robson. Specifically, for simplicity let us assume that $r(\Gamma) = \Gamma$, and so $B = \mathbb{Q}_q[\Gamma, \Gamma^{-1}]$. Let C denote a matrix for the action of ∇ on our basis (4.1). So C is a $4g+1 \times 4g+1$ matrix over B, and ∇ acts on elements as $\frac{d}{d\Gamma} + C$. Let W_j be the space consisting of vectors in $H^1_{dR}(\mathcal{X}/\mathcal{S})$ all of whose entries are $\mathbb{Q}_q$-linear combinations of Γ^i for $-j \le i \le j$. For c suitably large, ∇ maps W_j into W_{j+c}. Let K_j and C_j be the kernel and cokernel of $\nabla : W_j \to W_{j+c}$. Then

$$\dim C_j = (\dim W_{j+c} - \dim W_j) + \dim K_j \le 2c(4g+1) + (4g+1).$$

(Here the bound $\dim K_j \le 4g+1$ is obtained by considering local expansions around $T = \Gamma - 1$, say, and using the fact that the kernel of $\frac{d}{dT} + C$ has dimension at most $4g+1$.) As this bound is independent of j, the cokernel of ∇ also has dimension at most $(2c+1)(4g+1)$. To handle the general case, replace Γ by $r(\Gamma)$ and use $r(\Gamma)$-adic expansions.

Thus in conclusion we have shown that $H^2_{dR}(\mathcal{X})$ is finite-dimensional by proving $H^1_{dR}(\mathcal{X}/\mathcal{S})$ is a free B-module of rank $4g+1$, showing then that $\operatorname{coker}(\nabla)$ is finite dimensional, and using the isomorphism $\operatorname{coker}(\nabla) \cong H^2_{dR}(\mathcal{X})$.

5 What about Frobenius?

5.1 Overconvergent functions

In the previous section, we used the fibration to prove the finite dimensionality of $H^2_{dR}(\mathcal{X})$. The problem is that although the map $A \mapsto H^2_{dR}(\mathcal{X})$ is a covariant

functor, the map $\bar{A} \mapsto A$ is not. Specifically, the qth power map acts on the ring $\bar{A}$; however, there is no ring endomorphism of A which "lifts" the qth power map on the residue ring $\bar{A}$. In other words, our construction fails to lift the *Frobenius* - without the Frobenius we can't have a cohomological formula for the zeta function!

To get around this we have to modify our lifting. This modification brings p-adic analysis into play, and delicate questions of convergence now make everything a lot more difficult. Here is what we do: First, the space $\mathcal{X}$ has more functions defined on it than just those in A. We have a p-adic norm on A, and so can take p-adic limits of functions in A. Precisely, the p-adic completion $\hat{A}$ of A is a much more appropriate ring with which to work. Indeed, there is a lifting of the Frobenius map in this larger ring. Unfortunately, replacing A by $\hat{A}$ in the construction in Section 4 would give infinite dimensional cohomology spaces. We instead choose a slightly smaller ring $A^\dagger$.

Explicitly, let

$$B^\dagger := \left\{ \sum_{n=-\infty}^{\infty} \frac{b_n(\Gamma)}{r(\Gamma)^n} \,|\, b_n \in \mathbb{Q}_q[\Gamma],\, \deg(b_n) < \deg(r), \liminf(\mathrm{ord}(b_n)/|n|) > 0 \right\}.$$

This is the subring of functions in $\hat{B}$ which converge on a slightly larger open set than just the base space $\mathcal{S}$ itself. Let

$$A^\dagger \subset \left\{ \sum_{m=-\infty}^{\infty} \sum_{i=0}^{2g} \frac{a_{m,i}(\Gamma) X^i}{\sqrt{Q}^m} \,|\, a_{m.i} \in B^\dagger \right\}. \tag{5.1}$$

be the subring with the following decay conditions:

$$a_{m,i} = \sum_{n=-\infty}^{\infty} \frac{b_{m,i,n}(\Gamma)}{r(\Gamma)^n} \text{ with } \liminf(\mathrm{ord}(b_{m,i,n})/(|m|+|n|)) > 0 \text{ as } |m|+|n| \to \infty.$$

This is the subring of functions in $\hat{A}$ which converge on a slightly larger open set than just $\mathcal{X}$ itself. The rings $A^\dagger$ and $B^\dagger$ are called the *weak* or *dagger* completions of A and B. Their elements are called *overconvergent functions*. The reason for considering such functions is that if a series in $\hat{B}$, say, is the derivative of a similar looking series, then it may be that this similar looking series does not lie in $\hat{B}$. In other words, the ring $\hat{B}$ is a not closed under the "formal integration" of functions. (For example, $\sum_{n=0}^{\infty} p^n \Gamma^{p^n-1} \in \hat{B}$ but $\sum_{n=0}^{\infty} \Gamma^{p^n} \notin \hat{B}$.) Put another way, integrating a function which converges on $\mathcal{S}$ might give one that only converges on a relatively open proper subset of $\mathcal{S}$. The solution is to begin with functions that converge a little beyond $\mathcal{S}$; if they can be integrated then the integral still converges a little beyond $\mathcal{S}$. It turns out that this restriction is enough to ensure the cohomology spaces we construct are finite dimensional — some indication as to why will be given in Section 5.4

5.2 Lifting Frobenius

We can do this by first defining $\mathrm{Frob}_q(X) := X^q$, $\mathrm{Frob}_q(\Gamma) := \Gamma^q$, and $\mathrm{Frob}_q(c) = c^\sigma$ for $c \in \mathbb{Q}_q$ where σ is the automorphism of $\mathbb{C}_p$ lifting the qth power Frobenius automorphism on $\mathbb{F}_q$. Now Frob_q can be defined by continuity on elements in $A^\dagger$ provided we can work out where it sends $r(\Gamma)^{-1}$ and $\sqrt{Q}$. Certainly $\mathrm{Frob}_q(r(\Gamma)^{-1}) = 1/r^\sigma(\Gamma^q)$ where the map σ acts on the coefficients. We need to write this as an element in $B^\dagger$. Since $p|(r(\Gamma)^q - r^\sigma(\Gamma^q))$ we have

$$\frac{1}{r^\sigma(\Gamma^q)} = \frac{1}{r^q}\left(1 - p\frac{s}{r^q}\right)^{-1}$$

for some $s \in \mathbb{Q}_q[\Gamma]$ of degree at most $q\deg(r)$. Using the binomial expansion we can expand this to give an element in $B^\dagger$. Similarly, we must have that $\mathrm{Frob}_q(\sqrt{Q})^2 = \mathrm{Frob}_q(Q) = Q^\sigma(X^q, \Gamma^q)$. Defining

$$\mathrm{Frob}_q(\sqrt{Q}) := Q^{q/2}\left(1 - \frac{Q^q - Q^\sigma(X^q, \Gamma^q)}{Q^q}\right)^{1/2} \tag{5.2}$$

does the trick. The righthand-side squares to $Q^\sigma(X^q, \Gamma^q)$ and since $p|(Q(X)^q - Q^\sigma(X^q, \Gamma^q))$ it can be expanded as a series in $A^\dagger$.

5.3 Overconvergent F-isocrystals

Now we can go back through Section 4 and replace A and B by $A^\dagger$ and $B^\dagger$ whenever they occur. Also, we must insist that we only consider derivations which are continuous with respect to the p-adic norm. The spaces which we denoted $H^*_{dR}(\mathcal{X})$ and $H^*_{dR}(\mathcal{X}/\mathcal{S})$ should now be written $H^*_{MW}(\bar{\mathcal{X}})$ and $H^*_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$. These are the *Monsky-Washniter cohomology* (a.k.a rigid cohomology) spaces of our surface and our family of curves, respectively. The big difference is that we can now act on all our commutative diagrams by Frob_q. Specifically, functoriality forces $\mathrm{Frob}_q(dX) = qX^{q-1}dX$ and $\mathrm{Frob}_q(d\Gamma) = q\Gamma^{q-1}d\Gamma$. The map Frob_q now acts on the "daggered" version of (4.3) going "into the page", and one gets a "commutative cube" since Frob_q commutes with the derivation maps. The map Frob_q then descends to a map from the cokernels of the two horizontal arrows "into the page". There is already a vertical map, which we shall still call ∇, between these two cokernels, and so ones ends up with a new commutative diagram:

$$\begin{array}{ccccccc} 0 & \longrightarrow & H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}}) & \xrightarrow{\nabla} & H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})d\Gamma & \longrightarrow & 0 \\ & & \downarrow \mathrm{Frob}_q & & \downarrow \mathrm{Frob}_q & & \\ 0 & \longrightarrow & H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}}) & \xrightarrow{\nabla} & H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})d\Gamma & \longrightarrow & 0. \end{array} \tag{5.3}$$

We say that $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$ admits a *commuting connection and Frobenius map.* We saw that $H^1_{dR}(\mathcal{X}/\mathcal{S})$ was a free B-module of finite rank with basis the forms (4.1). If $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$ is a free $B^\dagger$-module on the same basis, then (5.3) defines an *overconvergent F-isocrystal on the base space $\bar{\mathcal{S}}$.* Specifically, an

overconvergent F-isocrystal on $\bar{\mathcal{S}}$ is a finitely generated locally free $B^\dagger$-module with a commuting connection and Frobenius map. (Of course, free of finite rank is a nice special case of finitely generated and locally free.) The ring $B^\dagger$ itself gives the "trivial" rank one example of an overconvergent F-isocrystal on $\bar{\mathcal{S}}$.

By analogy with Section 4.3, we can show that $H^2_{MW}(\bar{\mathcal{X}})$ is finite dimensional provided we can establish first that $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$ is indeed free of finite rank, and then that $\operatorname{coker}(\nabla)$ is finite dimensional. If the first space is indeed of finite rank, the latter space is called the *first cohomology space* of our overconvergent F-isocrystal $(H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}}), \nabla, \mathrm{Frob}_q)$.

5.4 Local study around missing points

Let us first consider $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$. Since $H^1_{dR}(\mathcal{X}/\mathcal{S})$ is spanned by the forms (4.1) (in fact they form a basis), one might hope the same is true for $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$. This is true, but it is quite surprising. The reason it is surprising is that as one reduces forms in AdX divisions occur; for example, reducing a form with $Q^{m/2+1}$ on the denominator to one with $Q^{m/2}$ requires division by m, see (4.2). Division will eventually introduce powers of the characteristic p on the denominator, and thus the form gets p-adically "larger and larger" as one reduces it. This suggests that if one takes a limit of such forms, i.e. and element in $A^\dagger dX$, it will reduce to a limit of "larger and larger" forms, and these limits might not always exist! However, some miraculous cancellation takes place, and the limits always do exist when one reduces forms in $A^\dagger dX$ (though not forms in the larger module $\hat{A}dX$).

To see what is really going on, one needs to study the reduction of forms "around the missing points"; this idea is due originally to Monsky [17]. Specifically, assume $U \in \mathbb{Q}_q[X, \Gamma]$ has *p-adic integral* coefficients, i.e., they all have p-adic order ≥ 0. Suppose we have iterated (4.2) $\lceil m/2 \rceil - 1$ times to obtain a relation

$$\frac{U dX}{\sqrt{Q}^m} - \frac{V dX}{Q} = \partial \left(\sum_{i\, even; 2 \leq i < m} \frac{W_{m-i}}{\sqrt{Q}^{m-i}} \right). \tag{5.4}$$

Here each $W_{m-i} \in \mathbb{Q}_q[\Gamma, r(\Gamma)^{-1}][X]$ has degree in X at most $2g$. A naive analysis shows that V becomes integral upon multiplication by $(m-2)(m-4)\ldots$. Let $n = p^c$ where $c = \max_j\{\mathrm{ord}(m-2j)\}$ and the max runs over positive j with $m - 2j > 0$. Notice that n/i is a p-adic integer for all positive rational numbers $i = m/2-1, m/2-2, \ldots$. We shall show that the form VdX/Q actually becomes integral upon multiplication by n. Since certainly $c \leq \log_p(m-2)$ this means that only "logarithmically small" powers of p are introduced in the denominator during reduction. (By contrast, the naive analysis suggests that the powers grow linearly during reduction!)

First, specialise $\Gamma = \gamma$ where $\gamma \in \mathbb{O}_p$ with $r(\gamma) \neq 0 \bmod p$. Let $a_1, \ldots, a_{2g+1} \in \mathbb{C}_p$ be the roots of $Q(X, \gamma)$; they are distinct modulo p. Take local expansions in terms of $T := X - a_1$. For example, the polynomial $\sqrt{Q}$ can be expanded as

$T^{1/2}\sum_{i=0}^{\infty} A_i T^i$ for some integral elements $A_i \in \mathbb{O}_p$. From (5.4) we get

$$T^{-m/2}\sum_{i=0}^{\infty} u_i T^i dT - T^{-1}\sum_{i=0}^{\infty} v_i T^i dT = \frac{d}{dT}\left(T^{-\frac{m}{2}+1}\sum_{i=0}^{\infty} w_i T^i\right). \qquad (5.5)$$

The leading coefficient w_0 on the righthand side is $W_{m-2}(a_1,\gamma)(\prod_{i\neq 1}\sqrt{a_1-a_i})^{-1}$. (Here the squareroot is the one which is chosen when expanding $(X-a_i)^{1/2} = (T+(a_1-a_i))^{1/2}$ as a series in T.) Notice that the second factor here is a p-adic unit, so has order zero. All of the elements u_i are integral, since U was assumed to have integral coefficients. Integrating (5.5) and comparing leading coefficients we see that $W_{m-2}(a_1,\gamma)$ is integral upon multiplication by $-m/2+1$; thus it is integral upon multiplication by n. Since this is true for all $2g+1$ roots, it follows that $nW_{m-2}(X,\gamma)$ itself must have integral coefficients. We can now subtract the integral of the local expansion of $W_{m-2}(X,\gamma)/\sqrt{Q(X,\gamma)}^{m-2}$ from both sides of the integrated version of (5.5). Now compare leading coefficients and deduce $(-m/2+2)W_{m-4}$ is integral etc. One concludes that nW_{m-i} is integral for all even i with $2 \leq i < m$ and hence that $nV(X,\gamma)$ is integral. Since this is true for all $\gamma \in \mathbb{O}_p$ with $r(\gamma) \neq 0 \bmod p$ it follows that $nV(X,\Gamma)$ is integral.

A similar argument looking at "local expansions around the missing point at infinity" handles the reduction of the degree in X to write VdX/Q as a linear combination of (4.1). Overall, the "logarithmically small" powers of p which are introduced on the denominator during reduction are swamped in the limit by the "overconvergence" of the series being reduced. Thus (4.1) also spans $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$. Again, it is actually a basis, although we will not prove this.

Having seen that $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$ is free of finite rank, we now turn our attention to coker(∇). We wish to show this is a finite dimensional $\mathbb{Q}_q$-vector space. Proving this is actually the central difficulty in p-adic cohomology. Indeed, Monsky comments in [19]: "the sticking point to proving finite dimensionality [of p-adic cohomology] seems to be ... the question of the finiteness of the cokernel for certain ordinary differential operators on rings of p-adic analytic functions". The author is not qualified to comment much on this problem, beyond saying that the solution lies in an understanding of the local structure of differential operators around missing points, and an application of the "localisation method" above [10, Section 6.4]. The local structure of such operators is described by the "p-adic local monodromy theorem" (a.k.a Crew's conjecture), which was proved independently by André [1], Kedlaya [9], and Mebkhout [16]. (We note that [5, Section 6(c)] seems to contain the first explicit study of this problem.) We refer the reader to the cited papers for more details on this problem; unfortunately, we will not give a proof of the finiteness of coker(∇).

6 A finiteness theorem with a sketch proof

In the previous three sections we have seen a proof that the cohomology space $H^2_{MW}(\bar{\mathcal{X}})$ of our surface is finite-dimensional (admittedly omitting some tech-

nical details in Section 5.4). The proof involved a number of steps: Fibre the surface into smooth curves; Prove the relative first cohomology of this family was of finite rank and thus defined an overconvergent F-isocrystal on the base; Show that the first cohomology of this overconvergent F-isocrystal was finite dimensional; Have a "spectral sequence" which compares $H^2_{MW}(\bar{\mathcal{X}})$ with this first cohomology space. Thus we reduced our problem for the two dimensional surface $\bar{\mathcal{X}}$ to that of showing finiteness of cohomology of an overconvergent F-isocrystal on the curve $\bar{\mathcal{S}}$. We could handle this case by a careful local study (that was the difficult bit we omitted). With a bit more cohomological machinery, we can construct an argument for the general case based upon this. Here it is:

Theorem 6.1. *The rigid cohomology of an overconvergent F-isocrystal defined on a smooth affine variety is finite dimensional.*

Theorem 2.2 is the special case where the overconvergent F-isocrystal is "trivial". We give a sketch-proof. It is somewhat idealised, and the real proof [10] takes a slightly different approach to circumvent some technical difficulties.

Proof. Our proof will be by induction on the dimension of the smooth affine variety. The case of a curve can be handled using the local techniques in Section 5.4, so we shall assume it is true in this case. Suppose now that the smooth affine variety $\bar{\mathcal{X}}$ is of dimension $n > 1$. For simplicity, let us suppose the F-isocrystal defined upon it is the "trivial one" — this just means that the cohomology we want to compute is that of the variety itself. Fibre $\bar{\mathcal{X}}$ into curves over affine space $\bar{\mathcal{S}}$ of dimension $n-1$. Unfortunately not all of the curves need be smooth. Let $\bar{\mathcal{X}}_0 \to \bar{\mathcal{S}}_0$ be the subfamily of smooth curves, where $\bar{\mathcal{X}}_0 \subseteq \bar{\mathcal{X}}$ and $\bar{\mathcal{S}}_0 \subseteq \bar{\mathcal{S}}$. Then $\bar{\mathcal{X}}_0$ is dense in $\bar{\mathcal{X}}$, and so the difference $\bar{\mathcal{X}} - \bar{\mathcal{X}}_0$ has dimension less than n. By induction we can assume its cohomology is finite dimensional. There is an exact sequence relating $H^*_{MW}(\bar{\mathcal{X}})$, $H^*_{MW}(\bar{\mathcal{X}} - \bar{\mathcal{X}}_0)$ and $H^*_{MW}(\bar{\mathcal{X}}_0)$ and the finite-dimensionality of the first follows from that of the second and third. Thus it is enough to prove finiteness for $H^*_{MW}(\bar{\mathcal{X}}_0)$. We have a fibration $\bar{\mathcal{X}}_0 \to \bar{\mathcal{S}}_0$ into smooth curves, exactly as in our example. The relative rigid cohomology $H^i_{MW}(\bar{\mathcal{X}}_0/\bar{\mathcal{S}}_0)$ for $i = 0$ and 1 define overconvergent F-isocrystals on $\bar{\mathcal{S}}_0$ - the difficult part is showing for $i = 1$ that it is (locally) free of *finite rank*, which can be done using a local argument. By the "Leray spectral sequence" for rigid cohomology, we can deduce the finiteness of $H^*_{MW}(\bar{\mathcal{X}}_0)$ from the finiteness of the cohomology of these two overconvergent F-isocrystals. Since $\dim(\bar{\mathcal{S}}_0) < n$ this can be assumed by induction. This completes the induction step. □

Note that if we had started out honestly with a general overconvergent F-isocrystal on $\bar{\mathcal{X}}$, we would have needed a *push forward* construction to push it down to one on $\bar{\mathcal{S}}$. Our relative construction is a special case of this. In the real proof [10], one restricts the overconvergent F-isocrystal to some dense open subset of $\bar{\mathcal{X}}$. This subset is chosen to be an unramified cover of affine space of dimension n. One pushes forward the restricted overconvergent F-isocrystal to this affine space, and then down to affine space of dimension one less. Then induction can be applied; see [12] for a more detailed overview.

7 Dwork's conjecture

We now turn our attention to Theorem 1.2. Again we shall try and explain its meaning by dint of our example. This section will use the notation introduced in Sections 3, 4 and 5.

Let $\bar{\gamma} \in \bar{\mathbb{F}}_q$ with $\bar{r}(\bar{\gamma}) \neq 0$. Write $\deg(\bar{\gamma})$ for the degree of the extension $\mathbb{F}_q(\bar{\gamma})/\mathbb{F}_q$. Then the fibre $\bar{\mathcal{X}}_{\bar{\gamma}}$ is a smooth curve defined over $\mathbb{F}_q(\gamma) = \mathbb{F}_{q^{\deg(\bar{\gamma})}}$ (see Section 3). We shall explain the meaning of Theorem 1.2 (Dwork's conjecture) in the case of our family $\bar{\mathcal{X}} \to \bar{\mathcal{S}}$ of curves $\bar{\mathcal{X}}_{\bar{\gamma}}$.

7.1 The cohomology of a fibre

We first need to understand the Monsky-Washnitzer cohomology spaces $H^*_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}})$. These are defined by lifting the coordinate ring of $\bar{\mathcal{X}}_{\bar{\gamma}}$, taking its dagger completion, and the homology of the corresponding de Rham complex. However, rather than go through all this again, we will just "specialise" the relative constructions in Sections 4 and 5. Specifically, for $i = 0, 1$ we have $H^i_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}}) = H^i_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}}) \otimes \mathbb{Q}_q(\gamma)$ with the tensor product via the specialisation map $\Gamma \to \gamma$. Here $\gamma \in \mathbb{O}_p$ is the unique element (Teichmüller lift) which reduces to $\bar{\gamma}$ modulo p such that $\gamma^{q^{\deg \bar{\gamma}}} = \gamma$. Moreover, the action of Frob_q commutes with this specialisation. In concrete terms, this just means that $H^1_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}})$ has as a $\mathbb{Q}_q(\gamma)$-basis the forms (4.1), with $Q(X, \Gamma)$ specialised to $Q(X, \gamma)$. Likewise, $H^0_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}})$ is the space of constants $\mathbb{Q}_q(\gamma)$, whereas $H^0_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$ was the ring $B^\dagger$ of elements in $A^\dagger$ killed by differentiation with respect to X. To see why all this is true, think about how we actually constructed the relative spaces: We took Γ to be a parameter; however, in the relative construction it could equally well have been a field element since we never took its derivative. Note though that $\mathrm{Frob}_q : \Gamma \mapsto \Gamma^q$ and so it only makes sense to specialise Γ in the construction to an element γ such that $\sigma : \gamma \to \gamma^q$. The Teichmüller liftings are unique with this property.

The benefit of constructing $H^*_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}})$ by specialising the module $H^*_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$ is that it allows us to study the Frobenius maps simultaneously for all the fibres in the family. Specifically, let us focus on the Frobenius map acting on $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$. This space is a free $B^\dagger$-module of rank $4g + 1$, and Frob_q is an additive map on this space. It is actually semi-linear, since $\mathrm{Frob}_q(bm) = b^\sigma(\Gamma^q)\mathrm{Frob}_q(m)$ for $b \in B^\dagger$ and $m \in H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$. In any case, its action is unique determined by that on the basis (4.1), and this can be described by a $4g + 1 \times 4g + 1$ matrix over $B^\dagger$. Let us write $(\mathrm{Frob}_q(\Gamma))$ for this matrix. For $\bar{\gamma} \in \mathbb{F}_q$ the qth power Frobenius action on $H^1_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}})$ is given by specialising the matrix $(\mathrm{Frob}_q(\Gamma))$ at $\Gamma = \gamma$. More generally, for $\gamma \in \bar{\mathbb{F}}_q$ we need to specialise a matrix for $\mathrm{Frob}_q^{\deg(\bar{\gamma})}$, since we are interested in the $q^{\deg(\bar{\gamma})}$th power map "$\mathrm{Frob}_{q^{\deg(\bar{\gamma})}}$" acting on cohomology. Since everything is semilinear, the matrix for this is actually $(\mathrm{Frob}_q(\Gamma))(\mathrm{Frob}_q^\sigma(\Gamma^q) \dots (\mathrm{Frob}_q^{\sigma^{\deg(\bar{\gamma})-1}}(\Gamma^{q^{\deg(\bar{\gamma})-1}}))$. Here σ acts on the coefficients of the entries in the matrices.

The Monsky cohomological formula (2.1) in this case tells us that

$$Z(\bar{\mathcal{X}}_{\bar{\gamma}}, T) = \frac{\det(1 - Tq^{\deg(\bar{\gamma})}\mathrm{Frob}_{q^{\deg(\bar{\gamma})}}^{-1} | H^1_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}}))}{1 - q^{\deg(\bar{\gamma})}T}.$$

Regarding the denominator, note that $H^0_{MW}(\bar{\mathcal{X}}_{\bar{\gamma}}) \cong \mathbb{Q}_q(\gamma)$ and $q^{\deg(\bar{\gamma})}\mathrm{Frob}_{q^{\deg(\bar{\gamma})}}^{-1}$ acts on it by multiplication by $q^{\deg(\bar{\gamma})}$.

7.2 Factorisation via eigenspaces

Going back to the relative construction, notice that the involution $\sqrt{\bar{Q}} \to -\sqrt{\bar{Q}}$ on $\bar{A}$ by functorality defines an involution on $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})$. Looking at the basis forms (4.1) we see that it splits it into negative and positive eigenspaces of dimensions $2g$ and $2g+1$ respectively. Explicitly, the negative eigenspace has basis the forms in (4.1) with $j = 1$, and the positive eigenspace has basis the forms with $j = 2$. Both the qth power map and $\frac{\partial}{\partial \Gamma}$ on $\bar{A}$ commute with the involution, and thus the two eigenspaces are stable under Frob_q and ∇.

This decomposition of the cohomology space when specialised at a fibre shows us that the numerator of $Z(\bar{\mathcal{X}}_{\bar{\gamma}}, T)$ factorises as $P_{\bar{\gamma}}(T)Q_{\bar{\gamma}}(T)$ where $P_{\bar{\gamma}}(T)$ is the reverse characteristic polynomial of $q^{\deg(\bar{\gamma})}\mathrm{Frob}_{q^{\deg(\bar{\gamma})}}^{-1}$ acting on the negative eigenspace, and $Q_{\bar{\gamma}}(T)$ that for the positive eigenspace. We have $\deg(P_{\bar{\gamma}}) = 2g$ and $\deg(Q_{\bar{\gamma}}) = 2g+1$ since our maps are clearly invertible. The polynomial $Q_{\bar{\gamma}}(T)$ is actually the inverse of the zeta function of the zero dimensional set consisting of the $2g+1$ points we removed from the affine curve, i.e., those with Y-coordinate zero [8, Section 3]. It is not difficult to prove that zeta functions of zero-dimensional sets are finite products of rational functions of the form $1/(1-T^d)$ for $d \geq 1$. Thus $Q_{\bar{\gamma}}(T)$ is a rather simple polynomial; in particular, its reciprocal roots are roots of unity. The functional equation for the zeta function of a smooth projective curve tells us that the reciprocal roots of $P_{\bar{\gamma}}(T)$ come in pairs which multiply together to give $q^{\deg(\bar{\gamma})}$. Thus $P_{\bar{\gamma}}(T)$ also equals the reverse characteristic polynomial of $\mathrm{Frob}_{q^{\deg(\bar{\gamma})}}$ itself on the negative eigenspace.

7.3 An L-function

Since Frob_q and ∇ commute with the involution, we can also decompose our overconvergent F-isocrystal $(H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}}), \nabla, \mathrm{Frob}_q)$ as the direct sum of one on the negative eigenspace and one on the positive eigenspace. Let's focus on the one on the negative eigenspace, since this gives the interesting part of the zeta function. We shall denote this $(H^1_{MW}(\bar{\mathcal{X}}, \bar{\mathcal{S}})_-, \nabla_-, \mathrm{Frob}_{q-})$. Concretely, the space $H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})_-$ is spanned by forms in (4.1) with $j = 1$. The action of Frob_{q-} on such a form can be calculated by first using the formula (5.2) and then reducing back to a linear combination of the basis elements using the algorithm in Section 4.2. Similarly, the action of ∇_- is given by differentiating basis elements with respect to Γ and then reducing.

Denote the $2g \times 2g$ matrix for Frob_{q-} as $(F(\Gamma))$. The discussions in the previous two subsections lead us to the equation

$$P_{\bar{\gamma}}(T) = \det(1 - (F(\gamma))(F(\gamma))^{\sigma} \dots (F(\gamma))^{\sigma^{\deg(\bar{\gamma})-1}} T).$$

The product $\prod_{\bar{\gamma}} P_{\bar{\gamma}}(T^{\deg(\bar{\gamma})})^{-1}$ over $\bar{\gamma} \in \bar{\mathbb{F}}_q$ with $\bar{r}(\bar{\gamma}) \neq 0$ is called the *L-function* attached to the overconvergent F-isocrystal $(H^1_{MW}(\bar{\mathcal{X}}/\bar{\mathcal{S}})_-, \nabla_-, \mathrm{Frob}_{q-})$. It is actually just the "interesting part" of the zeta function of our surface $\bar{\mathcal{X}}$, i.e., the part which does not come from the deleted curve $Y = \bar{Q}(X, \Gamma) = 0$, $\bar{r}(\Gamma) \neq 0$ or the deleted line at infinity. In particular, it is a rational function.

7.4 The unit root zeta function

A more mysterious function can be defined in the following way. Write $P_{\bar{\gamma}}(T) = \prod_{i=1}^{2g}(1 - \alpha_i T)$ and let $P^u_{\bar{\gamma}}(T)$ be the product of all factors $1 - \alpha_i T$ of $P_{\bar{\gamma}}(T)$ for which $\mathrm{ord}(\alpha_i) = 0$. It is the product over all roots which are units in the ring of integers of $\mathbb{C}_p$. It is known that $\deg(P^u_{\bar{\gamma}}) \leq g$, and when equality occurs we say that $\bar{\mathcal{X}}_{\bar{\gamma}}$ is *ordinary*. In fact, there is a polynomial $\bar{h}(\Gamma) \in \mathbb{F}_q[\Gamma]$ (Hasse polynomial) defined as the determinant of a $g \times g$ matrix such that $\bar{\mathcal{X}}_{\bar{\gamma}}$ is ordinary if and only if $\bar{h}(\bar{\gamma}) \neq 0$. Let us *assume* this polynomial is not identically zero, so all but finitely many fibres in the family are ordinary.

The *unit root zeta function* of the family $\bar{\mathcal{X}} \to \bar{\mathcal{S}}$ is the product $\prod_{\bar{\gamma}} P^u_{\bar{\gamma}}(T^{\deg(\bar{\gamma})})$. (More precisely, this is the unit root zeta function of the family in which we have replaced the ramification points.) Dwork conjectured that this is a p-adic meromorphic function, i.e., it can be written as a quotient of power series $a(T)/b(T)$ where each series converges on the whole of $\mathbb{C}_p$. This is the next best thing in the p-adic world to being a rational function. (It is known that the unit root zeta function is not rational in general; specifically, it is not rational for the universal family of elliptic curves.)

7.5 An idea on how it is proved

Dwork proved that the zeta function of a variety is rational by first showing it is p-adic meromorphic, and then applying an archimedean estimate to show it must be rational; see [13] for a nice exposition. Dwork's meromorphy proof extends without too much difficulty to a more general situation. Specifically, given a finite invertible matrix $(G(\Gamma))$ with entries in $B^{\dagger}$, one can attach an L-function $L(G, T)$ to it:

$$L(G, T) = \prod_{\bar{\gamma}} \det(1 - (G(\gamma))(G(\gamma))^{\sigma} \dots (G(\gamma))^{\sigma^{\deg(\bar{\gamma})-1}} T^{\deg(\bar{\gamma})})^{-1}.$$

Here the product is over $\bar{\gamma} \in \bar{\mathbb{F}}_q$ with $\bar{r}(\bar{\gamma}) \neq 0$. The simplest example is the 1×1 identity matrix. The L-function is then just the zeta function of $\bar{\mathcal{S}}$. Another example is $L(F, T)$, which is a rational factor in the zeta function of the surface $\bar{\mathcal{X}}$. Dwork's technique shows that $L(G, T)$ is always p-adic meromorphic [21].

(Note that when one can also find a commuting connection then this is the L-function of an overconvergent F-isocrystal. Theorem 6.1 and a generalisation of the Monsky cohomological formula (2.1) shows that the L-function is rational in this case. The function $L(F,T)$ is such an example.)

Removing any non-ordinary fibres in our family $\bar{\mathcal{X}} \to \bar{\mathcal{S}}$ gives a new family in which all of the fibres are ordinary. For simplicity, let us retain the notation $\bar{B}$ for the coordinate ring of the base of this new family. Each polynomial $P^u_{\bar{\gamma}}(T)$ which occurs in our ordinary family is now of degree g. One approach to proving Dwork's conjecture would be to find a $g \times g$ matrix $(F^u(\Gamma))$ over $B^\dagger$ such that

$$P^u_{\bar{\gamma}}(T) = \det(1 - (F^u(\gamma))(F^u(\gamma))^\sigma \dots (F^u(\gamma))^{\sigma^{\deg(\gamma)-1}} T)$$

for all $\bar{\gamma} \in \bar{\mathbb{F}}_q$ with $\bar{r}(\bar{\gamma}) \neq 0$. For then the unit root zeta function of our ordinary family would be the inverse of the L-function $L(F^u, T)$, and the technique of Dwork shows this is meromorphic. It turns out that it is possible to find a matrix over $\hat{B}$ with this property; however, unfortunately Dwork's method does not show that L-functions attached to such *convergent F-crystals* are meromorphic. Indeed, there is an example in which the L-function attached to a matrix over $\hat{B}$ is not meromorphic [21].

The proof of Dwork's conjecture [23, 24, 25] involves a sophisticated limiting argument that takes us out of the category of overconvergent F-isocrystals and into the larger category of (possibly) infinite rank modules over $B^\dagger$ with a Frobenius action. A generalisation of Dwork's meromorphy proof for L-functions $L(G,T)$ attached to infinite matrices $(G(\Gamma))$ over $B^\dagger$ then allows one to deduce the required results. Of course, this description is something of an oversimplification!

8 Attribution of results

The worked example in this paper is based on [8], extended to families of curves in the expository paper [14]. (The author's own work on the subject is on the problem of actually computing zeta functions of varieties over finite fields. For this it turns out that the relative construction is actually very useful, see [8, 14] for details and further references.) Theorem 1.1 was first proved independently by Grosse-Klönne [7] and Tsuzuki [20]. The special case of smooth affine varieties (Theorem 2.2) was proved earlier by Berthelot [2] and Mebkhout [16]. Theorem 6.1 is due to Kedlaya [10]; see [12] for an overview of the proof. (Note that Kedlaya's theorem is central to a proof of an analogue of "Deligne's Main Theorem" in the context of p-adic cohomology [11]; this includes a "p-adic proof" of the Riemann hypothesis for zeta functions of varieties over finite fields.) The proof of Dwork's conjecture is entirely due to Wan, and is contained in [23, 24, 25]; see also [22]. Dwork's conjecture itself was originally formulated in [6]

References

[1] Y. André, *Filtrations de type Hasse-Arf et monodromie p-adique*, Invent. Math. **148**, (2002), 285-317.

[2] P. Berthelot, *Finitude et pureté cohomologique rigide en cohomologie rigide*, Invent. Math. **128**, (1997), 80-124.

[3] D. Cox, J. Little, D. O'Shea, *Ideal, Varieties, and Algorithms*, 2nd Edition, Undergraduate Texts in Mathematics, Springer, 1997.

[4] B. Dwork, *On the rationality of the zeta function of an algebraic variety*, Amer. J. Math. **82**, (1960), 631-648.

[5] B. Dwork, *p-Adic cycles*, Pub. Math. IHES **37**, (1969), 27-115.

[6] B. Dwork, *Normalised period matrices II*, Ann. Math. **98**, (1973), 1-57.

[7] E. Grosse-Klönne, *Finiteness of de Rham cohomology in rigid analysis*, Duke. Math. J. **113**, (2002), 57-91.

[8] K. Kedlaya, *Counting points on hyperelliptic curves using Monsky-Washnitzer cohomology*, Journal of the Ramanujan Mathematical Society **16**, (2001), 323-338.

[9] K. Kedlaya, *A p-adic local monodromy theorem*, to appear in Ann. Math. Available at http://front.math.ucdavis.edu/math.AG/0110124.

[10] K. Kedlaya, *Finiteness of rigid cohomology with coefficients*, preprint. Version Oct 8, 2003.
Available at http://front.math.ucdavis.edu/math.AG/0208027.

[11] K. Kedlaya, *Fourier transforms and a p-adic "Weil II"*, preprint. Version March 2004.
Available at http://front.math.ucdavis.edu/math.NT/0210149.

[12] K. Kedlaya, *Crystals, Crew's conjecture, and cohomology*, Lecture notes available at: http://www-math.mit.edu/∼kedlaya/papers/

[13] N. Koblitz, *p-Adic numbers, p-adic analysis and zeta functions*, Graduate Texts in Mathematics 55, Springer, 1977.

[14] A.G.B. Lauder, *Rigid cohomology and p-adic point counting*, to appear in J. Théorie des Nombres de Bordeaux.
Available at http://www.maths.ox.ac.uk/∼lauder/

[15] Z. Mebkhout, *Sur le théorème de finitude de la cohomologie p-adique d'une variété affine non singulière*, Amer. J. Math. **119**, (1997), 1027-1081.

[16] Z. Mebkhout, *Analogue p-adique du Théorème de Turrittin et le Théorème de la monodromie p-adique*, Invent. Math. **148**, (2002), 319-351.

[17] P. Monsky, *One dimensional formal cohomology*, in Actes du Congrèss International des Mathématiciens (Nice 1970), Tome 1, 451-456, Gathier-Villars, Paris, 1971.

[18] P. Monsky, *Formal cohomology III: Fixed point theorems*, Ann. Math. **93**, (1971), 315-343.

[19] P. Monsky, *Finiteness of de Rham cohomology*, Amer. J. Math. **94**, (1972), 237-245.

[20] N. Tsuzuki, *Cohomological descent of rigid cohomology for proper coverings*, Invent. Math. **151**, (2003), 101-133.

[21] D. Wan, *Meromorphic continuation of L-functions of p-adic representations*, Ann. Math. **143**, (1996), 469-498.

[22] D. Wan, *A quick introduction to Dwork's conjecture*, Contemporary Mathematics **225**, (1999), 131-141.

[23] D. Wan, *Dwork's conjecture on unit root zeta functions*, Ann. Math., **150** (1999), 867-927.

[24] D. Wan, *Higher rank case of Dwork's conjecture*, J. Amer. Math. Soc. **13**, (2000), 807-852.

[25] D. Wan, *Rank one case of Dwork's conjecture*, J. Amer. Math. Soc. **13**, (2000), 853-908.