Annals of Mathematics, **163** (2006), 1019–1056

Dimers and amoebae

By Richard Kenyon, Andrei Okounkov, and Scott Sheffield*

Abstract

We study random surfaces which arise as height functions of random perfect matchings (a.k.a. dimer configurations) on a weighted, bipartite, doubly periodic graph G embedded in the plane. We derive explicit formulas for the surface tension and local Gibbs measure probabilities of these models. The answers involve a certain plane algebraic curve, which is the spectral curve of the Kasteleyn operator of the graph. For example, the surface tension is the Legendre dual of the Ronkin function of the spectral curve. The amoeba of the spectral curve represents the phase diagram of the dimer model. Further, we prove that the spectral curve of a dimer model is always a real curve of special type, namely it is a Harnack curve. This implies many qualitative and quantitative statement about the behavior of the dimer model, such as existence of smooth phases, decay rate of correlations, growth rate of height function fluctuations, etc.

Contents

*The third author was supported in part by NSF Grant No. DMS-0403182.

1. Introduction

A *perfect matching* of a graph is a collection of edges with the property that each vertex is incident to exactly one of these edges. A graph is *bipartite* if the vertices can be 2-colored, that is, colored black and white so that black vertices are adjacent only to white vertices and vice versa.

Random perfect matchings of a planar graph G—also called *dimer configurations*— are sampled uniformly (or alternatively, with a probability proportional to a product of the corresponding edge weights of G) from the set of all perfect matchings on G. These so-called dimer models are the subject of an extensive physics and mathematics literature. (See [9] for a survey.)

Since the set of perfect matchings of G is also in one-to-one correspondence with a class of height functions on the faces of G, we may think of random perfect matchings as (discretized) *random surfaces*. One reason for the interest in perfect matchings is that random surfaces of this type (and a more general class of random surfaces called *solid-on-solid models*) are popular models for crystal surfaces (e.g. partially dissolved salt crystals) at equilibrium. These height functions are most visually compelling when G is a honeycomb lattice. In this case, we may represent the vertices of G by triangles in a triangular lattice and edges of G by rhombi formed by two adjacent triangles. Dimer configurations correspond to tilings by such rhombi; they can be viewed as planar projections of surfaces of the kind seen in Figure 1. The third coordinate, which can be reconstructed from the dimer configuration uniquely, up to an overall additive constant, is the height function.

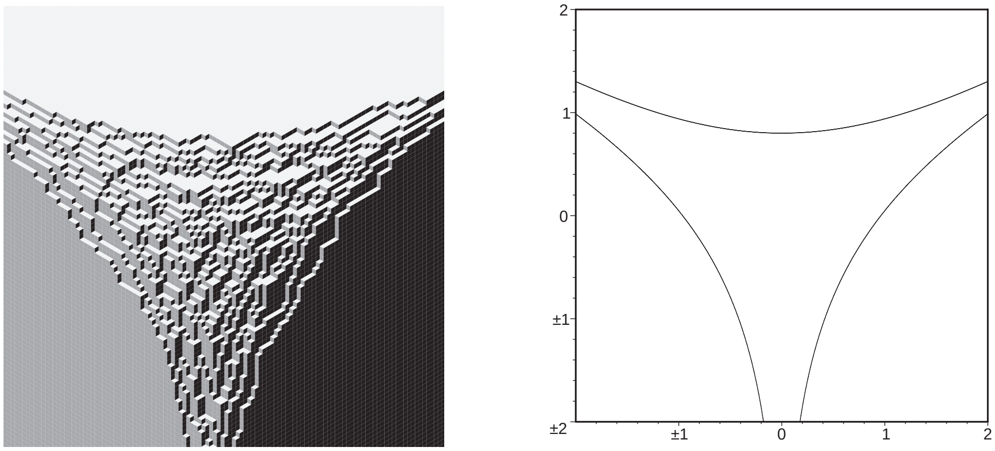

Figure 1. *On the left is the height function of a random volume-constrained dimer configuration on the honeycomb lattice. The boundary conditions here are that of a crystal corner: all dimers are aligned the same way deep enough in each of the three sectors. On the right is (the boundary of) the amoeba of a straight line.*

Most random surface models cannot be solved exactly, and we are content to prove qualitative results about the surface tension, the existence of facets, the set of gradient Gibbs measures, etc.

We will prove in this paper, however, that models based on perfect matchings (on any weighted doubly-periodic bipartite graph G in the plane) are exactly solvable in a rather strong sense. Not only can we derive explicit formulas for the surface tension—we also explicitly classify the set of Gibbs measures on tilings and explicitly compute the local probabilities in each of them. These results are a generalization of [2] where similar results for $G = \mathbb{Z}^2$ with constant edge weights were obtained.

In particular we show that Gibbs measures come in three distinct phases: a *rough*, or critical, phase, where the height fluctuations are on the order of $\log n$ for points separated by distance n, and correlations decay quadratically in n; a *frozen* phase where there are no large-scale fluctuations and the model is a Bernoulli process (points far apart are independent); and a *smooth* (sometimes referred to as rigid) phase where fluctuations have bounded variance, and correlations decay exponentially. We refer to these three phases respectively as liquid, frozen, and gaseous.

The theory has some surprising connections to algebraic geometry. In particular, in a sense described below, the phase diagram of dimer model on a weighted, doubly periodic graph (as one varies a two-parameter external magnetic field), is represented by the *amoeba* of an associated plane algebraic curve, the *spectral curve*; see Theorem 4.1. We recall that by definition [5], [14] the amoeba of an affine algebraic variety $X \in \mathbb{C}^n$ (plane curve, in our case) is the image of X under the map taking coordinates to the logarithms of their absolute value. See Figures 6 for an illustration of an amoeba with multiple holes. The so-called *Ronkin function* of the spectral curve (a function which is linear on each component of the complement of the amoeba and is strictly concave within the amoeba itself; see Figure 5) turns out to be the Legendre dual of the surface tension (Theorem 3.6).

Crystal facets in the model are in bijection with the components of the complement of the amoeba. In particular, the bounded ones correspond to compact holes in the amoeba; the number of bounded facets equals the genus of the spectral curve. By the Wulff construction, the Ronkin function describes the fine mesh limit height function of certain volume-constrained random surface models based on dimer height functions (which can in some cases be interpreted as the shape of a partially dissolved crystal corner). For example, the limit shape in the situation shown in Figure 1 is the Ronkin function of the straight line. It has genus zero and, hence, has no bounded facets. A more complicated limit shape, in which a bounded facet develops, can be seen in Figures 2 and 5.

Crystals that appear in nature typically have a small number of facets—the slopes of which are rational with respect to the underlying crystal lattice. But laboratory conditions have produced equilibrium surfaces with up to sixty identifiably different facet slopes [17]. It is therefore of interest to have a model

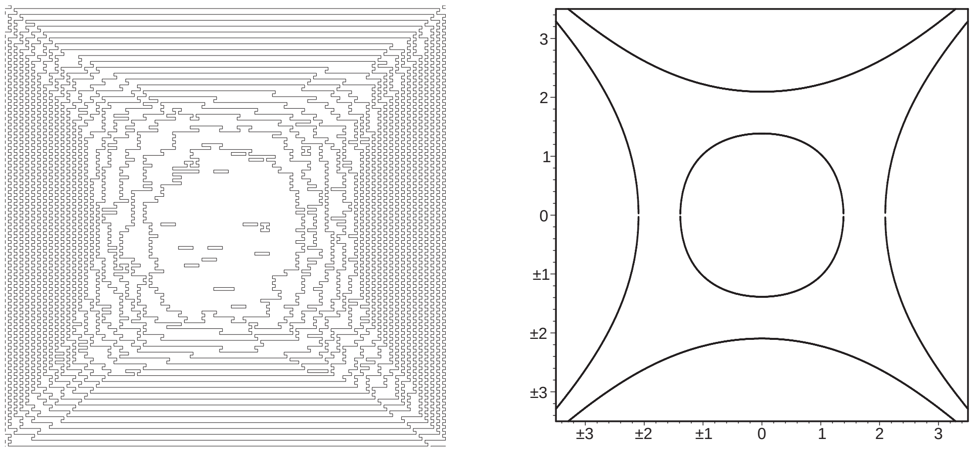

Figure 2. *On the left are the level sets of perimeter* 10 *or longer of the height function of a random volume-constrained dimer configuration on* $\mathbb{Z}^2$ (*with* 2×2 *fundamental domain*). *The height function is essentially constant in the middle — a facet is developing there. The intermediate region*, *in which the height function is not approximately linear*, *converges to the amoeba of the spectral curve*, *which can be seen on the right. The spectral curve in this case is a genus* 1 *curve with the equation* $z + z^{-1} + w + w^{-1} = 6.25$.

in which it is possible to generate crystal surfaces with arbitrarily many facets and to observe precisely how the facets evolve when weights and temperature are changed.

For another surprising connection between dimers and algebraic geometry see [16].

Acknowledgments. The paper was completed while R. K. was visiting Princeton University. A. O. was partially supported by DMS-0096246 and a fellowship from Packard foundation. S. S. was supported in part by NSF Grant No. DMS-0403182.

2. Definitions

2.1. *Combinatorics of dimers.*

2.1.1. *Periodic bipartite graphs and matchings.* Let G be a $\mathbb{Z}^2$-periodic bipartite planar graph. By this we mean G is embedded in the plane so that translations in $\mathbb{Z}^2$ act by color-preserving isomorphisms of G — isomorphisms which map black vertices to black vertices and white to white. An example of such a graph is the square-octagon graph, the fundamental domain of which is shown in Figure 3. More familiar (and, in a certain precise sense, universal [12]) examples are the standard square and honeycomb lattices. Let G_n be the quotient of G by the action of $n\mathbb{Z}^2$. It is a finite bipartite graph on a torus.

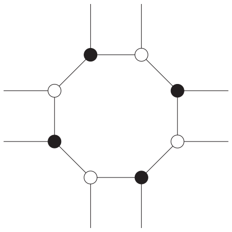

Figure 3. *The fundamental domain of the square-octagon graph.*

Let $\mathcal{M}(G)$ denote the set of perfect matchings of G. A well-known necessary and sufficient condition for the existence of a perfect matching of G is the existence of a unit flow from white vertices to black vertices, that is a flow with source 1 at each white vertex and sink 1 at every black vertex. If a unit flow on G exists, then by taking averages of the flow over larger and larger balls and subsequential limits one obtains a unit flow on G_1. Conversely, if a unit flow on G_1 exists, it can be extended periodically to G. Hence G_1 has a perfect matching if and only if G has a perfect matching.

2.1.2. *Height function.* Any matching M of G defines a white-to-black unit flow ω: flow by one along each matched edge. Let M_0 be a fixed periodic matching of G and ω_0 the corresponding flow. For any other matching M with flow ω, the difference $\omega - \omega_0$ is a divergence-free flow. Given two faces f_0, f_1 let γ be a path in the dual graph G^* from f_0 to f_1. The total flux of $\omega - \omega_0$ across γ is independent of γ and therefore is a function of f_1 called the *height function* of M.

The height function of a matching M is well-defined up to the choice of a base face f_0 and the choice of reference matching M_0. The difference of the height functions of two matchings is well-defined independently of M_0.

A matching M_1 of G_1 defines a periodic matching M of G; we say M_1 has *height change* (j, k) if the horizontal and vertical height changes of M for one period are j and k respectively, that is

$$h(v + (x, y)) = h(v) + jx + ky$$

where h is the height function on M. The height change is an element of $\mathbb{Z}^2$, and can be identified with the homology class in $H_1(\mathbb{T}^2, \mathbb{Z})$ of the flow $\omega_1 - \omega_0$. The height change of a larger graph G_n is defined analogously, replacing G_1 with G_n. (In particular, if M_1 is extended periodically to a matching M_n of G_n, then the height change of M_n is n times that of M_1.)

2.2. *Gibbs measures.*

2.2.1. *Definitions.* Let $\mathcal{E}$ be a real-valued function on edges of G_1, the *energy* of an edge. It defines a periodic energy function on edges of G. We define the energy of a finite set M of edges by $\mathcal{E}(M) = \sum_{e \in M} \mathcal{E}(e)$.

A *Gibbs measure* on $\mathcal{M}(G)$ is a probability measure with the following property. If we fix any finite subgraph of G, then conditioned on the edges that lie outside of G, the probability of any interior matching M of the remaining vertices is proportional to $e^{-\mathcal{E}(M)}$. An *ergodic Gibbs measure* (EGM) is a Gibbs measure on $\mathcal{M}(G)$ which is invariant and ergodic under the action of $\mathbb{Z}^2$.

For an EGM μ let $s = \mathbb{E}[h(v+(1,0))-h(v)]$ and $t = \mathbb{E}[h(v+(0,1))-h(v)]$ be the expected horizontal and vertical height change. We then have

$$\mathbb{E}[h(v+(x,y))-h(v)] = sx + ty.$$

We call (s,t) the *slope* of μ.

2.2.2. *Gibbs measures of fixed slope.* On $\mathcal{M}(G_n)$ we define a probability measure μ_n satisfying

$$\mu_n(M) = \frac{e^{-\mathcal{E}(M)}}{Z},$$

for any matching $M \in \mathcal{M}(G_n)$. Here Z is a normalizing constant known as the *partition function.*

For a fixed $(s,t) \in \mathbb{R}^2$, let $\mathcal{M}_{s,t}(G_n)$ be the set of matchings of G_n whose height change is $(\lfloor ns \rfloor, \lfloor nt \rfloor)$. Assuming that $\mathcal{M}_{s,t}(G_n)$ is nonempty, let $\mu_n(s,t)$ denote the conditional measure induced by μ_n on $\mathcal{M}_{s,t}(G_n)$.

The following results are found in Chapters 8 and 9 of [18]:

THEOREM 2.1 ([18]). *For each (s,t) for which $\mathcal{M}_{s,t}(G_n)$ is nonempty for n sufficiently large, $\mu_n(s,t)$ converges as $n \to \infty$ to an* EGM *$\mu(s,t)$ of slope (s,t). Furthermore μ_n itself converges to $\mu(s_0,t_0)$ where (s_0,t_0) is the limit of the slopes of μ_n. Finally, if (s_0,t_0) lies in the interior of the set of (s,t) for which $\mathcal{M}_{s,t}(G_n)$ is nonempty for n sufficiently large, then every* EGM *of slope (s,t) is of the form $\mu(s,t)$ for some (s,t) as above; that is, $\mu(s,t)$ is the unique* EGM *of slope (s,t).*

2.2.3. *Surface tension .* Let

$$Z_{s,t}(G_n) = \sum_{M \in \mathcal{M}_{s,t}(G_n)} e^{-\mathcal{E}(M)}$$

be the partition function of $\mathcal{M}_{s,t}(G_n)$. Define

$$Z_{s,t}(G) = \lim_{n\to\infty} Z_{s,t}(G_n)^{1/n^2}.$$

The existence of this limit is easily proved using subadditivity as in [2]. The function $Z_{s,t}(G)$ is the *partition function per fundamental domain* of $\mu(s,t)$ and

$$\sigma(s,t) = -\log Z_{s,t}(G)$$

is called the *surface tension* or *free energy* per fundamental domain. The explicit form of this function is obtained in Theorem 3.6.

The measure $\mu(s_0, t_0)$ in Theorem 2.1 above is the one which has minimal free energy per fundamental domain. Since the surface tension is strictly convex (see Chapter 8 of [18] or Theorem 3.7 below), the surface-tension minimizing slope is unique and equal to (s_0, t_0).

2.3. *Gauge equivalence and magnetic field.*

2.3.1. *Gauge transformations.* Since G is bipartite, each edge $e = (\mathsf{w}, \mathsf{b})$ has a natural orientation: from its white vertex w to its black vertex b. Any function f on the edges can therefore be canonically identified with a 1-form, that is, a function on oriented edges satisfying $f(-e) = -f(e)$, where $-e$ is the edge e with its opposite orientation. We will denote by $\Omega^1(G_1)$ the linear space of 1-forms on G_1. Similarly, Ω^0 and Ω^2 will denote functions on vertices and oriented faces, respectively.

The standard differentials

$$0 \to \Omega^0 \xrightarrow{d} \Omega^1 \xrightarrow{d} \Omega^2 \to 0$$

have the following concrete meaning in the dimer problem. Given two energy functions $\mathcal{E}_1$ and $\mathcal{E}_2$, we say that they are *gauge equivalent* if

$$\mathcal{E}_1 = \mathcal{E}_2 + df\,, \quad f \in \Omega^0\,,$$

which means that for every edge $e = (\mathsf{w}, \mathsf{b})$

$$\mathcal{E}_1(e) = \mathcal{E}_2(e) + f(\mathsf{b}) - f(\mathsf{w})\,,$$

where f is some function on the vertices. It is clear that for any perfect matching M, the difference $\mathcal{E}_1(M) - \mathcal{E}_2(M)$ is a constant independent of M, hence the energies $\mathcal{E}_1$ and $\mathcal{E}_2$ induce the same probability distributions on dimer configurations.

2.3.2. *Rotations along cycles.* Given an oriented cycle

$$\gamma = \{\mathsf{w}_0, \mathsf{b}_0, \mathsf{w}_1, \mathsf{b}_1, \dots, \mathsf{b}_{k-1}, \mathsf{w}_k\}\,, \quad \mathsf{w}_k = \mathsf{w}_0\,,$$

in the graph G_1, we define

$$\int_\gamma \mathcal{E} = \sum_{i=1}^{k-1} \left[\mathcal{E}(\mathsf{w}_i, \mathsf{b}_i) - \mathcal{E}(\mathsf{w}_{i+1}, \mathsf{b}_i)\right].$$

It is clear that $\mathcal{E}_1$ and $\mathcal{E}_2$ are gauge equivalent if and only if $\int_\gamma \mathcal{E}_1 = \int_\gamma \mathcal{E}_2$ for all cycles γ. We call $\int_\gamma \mathcal{E}$ the *magnetic flux* through γ. It measures the change in energy under the following basic transformation of dimer configurations.

Suppose that a dimer configuration M is such that every other edge of a cycle γ is included in M. Then we can form a new configuration M' by

$$M' = M \triangle \gamma\,,$$

where $\triangle$ denotes the symmetric difference. This operation is called *rotation along* γ. It is clear that

$$\mathcal{E}(M') = \mathcal{E}(M) \pm \int_\gamma \mathcal{E}\,.$$

The union of any two perfect matchings M_1 and M_2 is a collection of closed loops and one can obtain M_2 from M_1 by rotating along all these loops. Therefore, the magnetic fluxes uniquely determine the relative weights of all dimer configurations.

2.3.3. *Magnetic field coordinates.* Since the graph G_1 is embedded in the torus, the results of the previous section imply that the gauge equivalence classes of energies are parametrized by $\mathbb{R}^{F-1} \oplus \mathbb{R}^2$, where F is the number of faces of G_1. The first summand is $d\mathcal{E}$, a function on the faces subject to one relation: the sum is zero. We will denote the function $d\mathcal{E} \in \Omega^2(G_1)$ by B_z. We write $B := (B_x, B_y, B_z)$ where the other two parameters

$$(B_x, B_y) \in \mathbb{R}^2$$

are the magnetic flux along a cycle winding once horizontally (resp. vertically) around the torus.

In practice we will fix B_z and vary B_x, B_y, as follows. Let γ_x be a path in the dual of G_1 winding once horizontally around the torus. Suppose that k edges of G_1 are crossed by γ_x. On each edge of G_1 crossed by γ_x, add energy $\pm\frac{1}{k}\Delta B_x$ according to whether the upper vertex (the one to the left of γ_x when γ_x is oriented in the positive x-direction) is black or white. Similarly, let γ_y be a vertical path in the dual of G_1, crossing k' edges of G_1; add $\pm\frac{1}{k'}\Delta B_y$ to the energy of edges crossed by γ_y according to whether the left vertex is black or white.

The new magnetic field is now $B' = B + (0, \Delta B_x, \Delta B_y)$. This implies that the change in energy of a matching under this change in magnetic field depends linearly on the height change of the matching:

LEMMA 2.2. *For a matching M of G_1 with height change (h_x, h_y) we have*

$$\mathcal{E}_{B'}(M) - \mathcal{E}_{B'}(M_0) = \mathcal{E}_B(M) - \mathcal{E}_B(M_0) + \Delta B_x h_x + \Delta B_y h_y\,.$$

3. Surface tension

3.1. *Kasteleyn matrix and characteristic polynomial.*

3.1.1. *Kasteleyn weighting.* A Kasteleyn matrix for a finite bipartite *planar* graph Γ is a weighted, signed adjacency matrix for Γ, whose determinant is the partition function for matchings on Γ. It can be defined as follows. Multiply the edge weight of each edge of Γ by 1 or -1 in such a way that the following holds: around each face there are an odd number of $-$ signs if the face has $0 \bmod 4$ edges, and an even number if the face has $2 \bmod 4$ edges. This is always possible [8]. Although the Kasteleyn matrix is not uniquely determined, it is uniquely determined as a function of the edge weights once we choose these signs.

Let $K = (K_{\mathsf{wb}})$ be the matrix with rows indexed by the white vertices and columns indexed by the black vertices, with K_{wb} being the above signed edge weight $\pm e^{-\mathcal{E}((\mathsf{w},\mathsf{b}))}$ (and 0 if there is no edge). Kasteleyn proved [8] that $|\det K|$ is the partition function,

$$|\det K| = Z(\Gamma) = \sum_{m\in\mathcal{M}(\Gamma)} e^{-\mathcal{E}(m)}.$$

3.1.2. *Periodic boundary conditions.* For bipartite graphs embedded in a torus, one can construct a Kasteleyn matrix K as above [8]. As in the previous section, we assume that we have fixed the signs of the edges, so that the Kasteleyn matrix is determined as a function of the edge weights. Then $|\det K|$ is a signed sum of weights of matchings, where the sign of a matching depends on the parity of its horizontal and vertical height change. This sign is a function on $H_1(\mathbb{T}^2, \mathbb{Z}/2\mathbb{Z})$, that is, matchings with the same horizontal and vertical height change modulo 2 appear in $\det K$ with the same sign. Moreover of the four possibly parity classes, three have the same sign in $\det K$ and one has the opposite sign [19]. (For the reader familiar with this terminology, this sign function is one of the 4 *spin structures* or theta characteristics on the torus.)

The sign depends on the choices in the definition of the Kasteleyn matrix. By an appropriate choice we can make the $(0, 0)$ parity class (whose height changes are both even) have even sign and the remaining classes have odd sign, that is, $\det K = M_{00} - M_{10} - M_{01} - M_{11}$, where M_{00} is the partition function for matchings with even horizontal and vertical height changes, and so on.

The actual partition function can then be obtained as a sum of four determinants

$$Z = \frac{1}{2}(-Z^{(00)} + Z^{(10)} + Z^{(01)} + Z^{(11)}),$$

where $Z^{(\theta\tau)}$ is the determinant of K in which the signs along a horizontal dual cycle (edges crossing a horizontal path in the dual) have been multiplied by $(-1)^\theta$ and along a vertical cycle have been multiplied by $(-1)^\tau$. (Changing the signs along a horizontal dual cycle has the effect of negating the weight of matchings with odd horizontal height change, and similarly for vertical.) For details see [8], [19].

3.1.3. *Characteristic polynomial.* Let K be a Kasteleyn matrix for the graph G_1 as above. Given any positive parameters z and w, we construct a "magnetically altered" Kastelyn matrix $K(z,w)$ from K as follows.

Let γ_x and γ_y be the paths introduced in Section 2.3.3. Multiply each edge crossed by γ_x by $z^{\pm 1}$ depending on whether the black vertex is on the left or on the right, and similarly for γ_y. See Figure 4 for an illustration of this procedure in the case of the honeycomb graph with 3 × 3 fundamental domain. We will refer to $P(z,w) = \det K(z,w)$ as the *characteristic polynomial* of G. The description of $\det K(z,w)$ as a signed partition function above implies that up to reflections $z \to -z$ and $w \to -w$ of the inputs, $P(z,w)$ is independent of the choice of signs used in defining K.

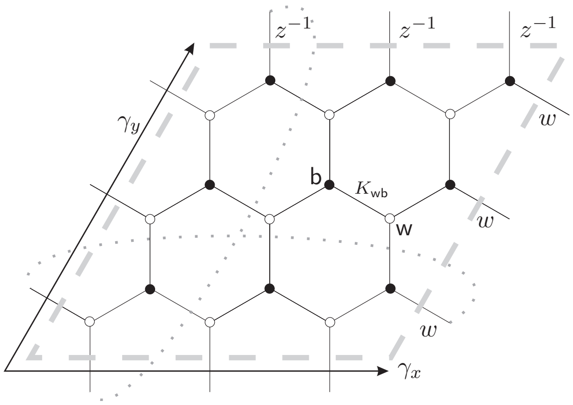

Figure 4. *The operator $K(z,w)$*

For example, for the square-octagon graph from Figure 3 this gives

$$P(z,w) = z + \frac{1}{z} + w + \frac{1}{w} + 5\,. \tag{1}$$

Recall that M_0 denotes the reference matching in the definition of the height function and ω_0 denotes the corresponding flow. Let x_0 denote the total flow of ω_0 across γ_x and similarly let y_0 the total flow of ω_0 across γ_y. The above remarks imply the following:

Proposition 3.1. *We have*

$$P(z,w) = z^{-x_0} w^{-y_0} \sum_{M \in \mathcal{M}(G_1)} e^{-\mathcal{E}(M)} z^{h_x} w^{h_y} (-1)^{h_x h_y + h_x + h_y}$$

where $h_x = h_x(M)$ and $h_y = h_y(M)$ are the (*integer*) *horizontal and vertical height change of the matching M and $\mathcal{E}(M)$ is its energy.*

Since G_1 has a finite number of matchings, $P(z,w)$ is a Laurent polynomial in z and w with real coefficients. The coefficients are negative or zero except when h_x and h_y are both even. Note that if the coefficients in the definition of P

were replaced with their absolute values (i.e., if we ignored the $(-1)^{h_x h_y + h_x + h_y}$ factor), then $P(1,1)$ would be simply the partition function $Z(G_1)$, and $P(z,w)$ (with z and w positive) would be the partition function obtained using the modified energy $\mathcal{E}'(M) = \mathcal{E}(M) + h_x \log z + h_y \log w$. With the signs, however, $P((-1)^\theta, (-1)^\tau) = Z^{(\theta\tau)}$ and the partition function may be expressed in terms of the characteristic polynomial as follows:

$$Z = \frac{1}{2}\left(-P(1,1) + P(1,-1) + P(-1,1) + P(-1,-1)\right).$$

As we will see, all large-scale properties of the dimer model depend only on the polynomial $P(z,w)$.

3.1.4. *Newton polygon and allowed slopes.* By definition, the Newton polygon $N(P)$ of P is the convex hull in $\mathbb{R}^2$ of the set of integer exponents of monomials in P, that is

$$N(P) = \text{convex hull}\left\{(j,k) \in \mathbb{Z}^2 \,\middle|\, z^j w^k \text{ is a monomial in } P(z,w)\right\}.$$

PROPOSITION 3.2. *The Newton polygon is the set of possible slopes of translation invariant measures, that is, there exists a translation invariant measure of slope (s,t) if and only if $(s,t) \in N(P)$.*

Proof. A translation-invariant measure of average slope (s,t) determines a unit white-to-black flow on G_1 with vertical flux s and horizontal flux t: the flow along an edge is the probability of that edge occurring. However the matchings of G_1 are the vertices of the polytope of unit white-to-black flows of G_1, and the height change (s,t) is a linear function on this polytope. Therefore (s,t) is contained in $N(P)$ by Proposition 3.1.

If M_1 is the matching corresponding to a vertex of $N(P)$, then the slope corresponding to that vertex is achieved by the the singleton measure in which the tiling is a periodic extension of M_1 with probability one. All interior slopes are given by measures that are weighted averages of these periodic ones. Now [18], Theorem 9.1.1 proves that there is an (in fact unique) EGM of slope (s,t) for every slope (s,t) for which there is a translation-invariant measure. □

Note that changing the reference matching M_0 in the definition of the height function merely translates the Newton polygon.

3.2. *Asymptotics.*

3.2.1. *Enlarging the fundamental domain.* Characteristic polynomials of larger graphs may be computed recursively as follows:

THEOREM 3.3. *Let P_n be the characteristic polynomial of G_n. Then*

$$P_n(z,w) = \prod_{z_0^n = z,} \prod_{w_0^n = w} P(z_0, w_0).$$

Proof. We follow the argument of [2] where this fact is proved for grid graphs. Since symmetry implies that the right side is a polynomial in z and w, it is enough to check this statement for positive values of z and w. View the Kastelyn matrix $K_n(z,w)$ of G_n as a linear map from the space V_w of functions on white vertices of G_n to the space V_b of functions on black vertices. When α and β are nth roots of unity, let $V_w^{\alpha,\beta}$ and $V_b^{\alpha,\beta}$ be the subspaces of functions for which translation by one period in the horizontal or vertical direction corresponds to multiplication by α and β respectively. (This spaces can also be defined using the discrete Fourier transform.) Clearly, these subspaces give orthogonal decompositions of V_w and V_b, and $K_n(z,w)$ is block diagonal in the sense that it sends an element in $V_w^{\alpha,\beta}$ to an element in $V_b^{\alpha,\beta}$. We may thus write $\det K_n(z,w)$ as a product of the determinants of the n^2 restricted linear maps from $V_w^{\alpha,\beta}$ to $V_b^{\alpha,\beta}$; these determinants are given by $\det K(\alpha z^{1/n}, \beta w^{1/n})$. □

This recurrence relation allows us to compute partition functions on general G_n in terms of P:

Corollary 3.4.

$$Z(G_n) = \frac{1}{2}(-Z_n^{(00)} + Z_n^{(01)} + Z_n^{(10)} + Z_n^{(11)}), \tag{2}$$

where

$$Z_n^{(\theta\tau)} = P_n((-1)^\theta, (-1)^\tau) = \prod_{z^n=(-1)^\theta,\, w^n=(-1)^\tau} \prod P(z,w). \tag{3}$$

3.2.2. *Partition function per fundamental domain.* We are interested in the asymptotics of $Z(G_n)$ when n is large. The logarithm of the expression (3) is a Riemann sum for an integral over the unit torus $\mathbb{T}^2 = \{(z,w) \in \mathbb{C}^2 \,:\, |z| = |w| = 1\}$ of $\log P$; thus

$$\frac{1}{n^2}\log Z_n^{(\theta\tau)} = \frac{1}{(2\pi i)^2}\int_{\mathbb{T}^2} \log|P(z,w)|\frac{dz}{z}\frac{dw}{w} + o(1)$$

on condition that none of the points

$$\{(z,w) : z^n = (-1)^\theta, w^n = (-1)^\tau\} \tag{4}$$

falls close to a zero of P. If it does, such a point will affect the sum only if it falls within $e^{-O(n^2)}$ of a zero of P (and in any case can only decrease the sum). In this case, for any n' near but not equal to n, no point of the form (4) will fall so close to this zero of P. In Theorem 5.1 below we prove that P has at most two simple zeros on the unit torus. It follows that only for a very rare set of n does the Riemann sum not approximate the integral.

Note that $Z(G_n) \geq Z_n^{(\theta,\tau)}$ since these both count all configurations but the latter has some $-$ signs. Since by (2), $Z(G_n)$ satisfies

$$Z_n^{(\theta\tau)} \leq Z(G_n) \leq 2 \max_{\theta,\tau}\{|Z_n^{(\theta,\tau)}|\},$$

we have

$$\lim_{n\to\infty}{}' \frac{1}{n^2} \log Z(G_n) = \frac{1}{(2\pi i)^2} \int_{\mathbb{T}^2} \log |P(z,w)| \frac{dz}{z}\frac{dw}{w},$$

where the $\lim'$ means that the limit holds except possibly for a rare set of ns. But now a standard subadditivity argument (see e.g. [2]) shows that $Z(G_n)^{1/n^2} \leq Z(G_m)^{1/m^2}(1+o(1))$ for all large m so that in fact the limit exists without having to take a subsequence.

Theorem 3.5. *We have*

$$\log Z \overset{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n^2} \log Z(G_n) = \frac{1}{(2\pi i)^2} \int_{\mathbb{T}^2} \log |P(z,w)| \frac{dz}{z}\frac{dw}{w}.$$

The quantity Z is the partition function per fundamental domain.

3.2.3. *The amoeba and Ronkin function of a polynomial.* Given a polynomial $P(z,w)$, its *Ronkin function* is by definition the following integral

$$F(x,y) = \frac{1}{(2\pi i)^2} \int_{\mathbb{T}^2} \log |P(e^x z, e^y w)| \frac{dz}{z}\frac{dw}{w}\,. \tag{5}$$

A closely related object is the *amoeba* of the polynomial P defined as the image of the curve $P(z,w)=0$ in $\mathbb{C}^2$ under the map

$$(z,w) \mapsto (\log|z|, \log|w|)\,.$$

We will call the curve $P(z,w)=0$ the *spectral curve* and denote its amoeba by $\mathbb{A}(P)$.

It is clear that the integral (5) is singular if and only if (x,y) lies in the amoeba. In fact, the Ronkin function is linear on each component of the amoeba complement and strictly convex over the interior of the amoeba (in particular, implying that each component of $\mathbb{R}^2 \setminus \mathbb{A}(P)$ is convex). This and many other useful facts about the amoebas and Ronkin function can be found in [14]. See Figures 5 and 6 for an illustration of these notions.

We distinguish between the unbounded complementary components and the bounded complementary components.

3.2.4. *Surface tension.* Theorem 2.1 gave, for a fixed magnetic field, a two-parameter family of EGMs $\{\mu(s,t)\}$. (No magnetic field was mentioned in Theorem 2.1, but since the result was for arbitrary edge weights, can modify the edge weights the same way they would be modified by a magnetic field.) Let us vary the magnetic field as in Section 2.3.3. Let $\tilde{\mu}_n(B_x,B_y)$ be the

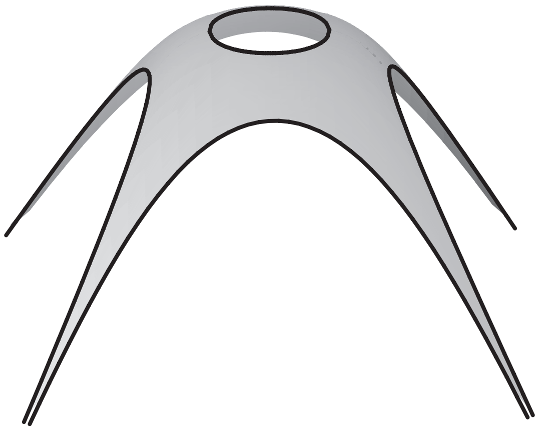

Figure 5. *The curved part of minus the Ronkin function of* $z + \frac{1}{z} + w + \frac{1}{w} + 5$. *This is the limit height function shape for square-octagon dimers with crystal corner boundary conditions.*

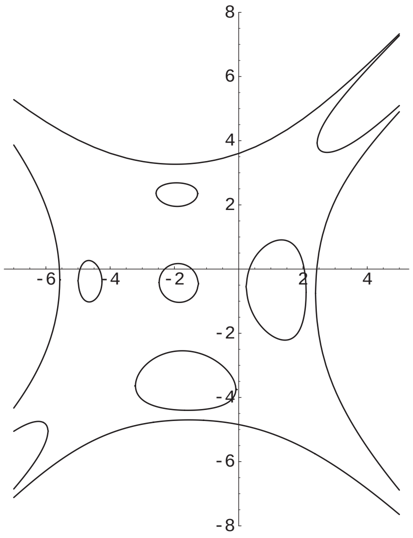

Figure 6. *The amoeba corresponding to the model of Figure 8. Its complement has one bounded component for each of the five interior integer points of* $N(P) = \{x : |x|_1 \leq 2\}$. *It also has two "semi-bounded" (i.e., contained in a strip of finite width) components, corresponding to two of the noncorner integer points on the boundary of* $N(P)$. *The four large components correspond to the corner vertices of* $N(P)$. *In the case of equal weights, the holes in the amoeba shrink to points and only four large unbounded components of the complement are present.*

measure on dimer configurations on G_n in the presence of an additional parallel magnetic field B_x, B_y. We define $\tilde{\mu}(B_x, B_y)$ to be the limit of $\tilde{\mu}_n(B_x, B_y)$ as $n \to \infty$ (which exists, by [18]) and let Z_{B_x,B_y} be its partition function per fundamental domain.

We can compute Z_{B_x,B_y} in two different ways. On the one hand, using Proposition 3.1, the characteristic polynomial becomes $P(e^{-B_x}z, e^{-B_y}w)$ and, hence, by Theorem 3.3 we have $Z_{B_x,B_y} = F(B_x, B_y)$, where F is the the Ronkin function of P. On the other hand, using Lemma 2.2 and basic properties of the surface tension (see Section 2.2.3) we obtain

$$F(B_x, B_y) = \max_{(s,t)} \left(-\sigma(s,t) + sB_x + tB_y\right) . \tag{6}$$

In other words, F is the Legendre dual of the surface tension. Since the surface tension is strictly convex, the Legendre transform is involutive and we obtain the following

THEOREM 3.6. *The surface tension $\sigma(s,t)$ is the Legendre transform of the Ronkin function of the characteristic polynomial P.*

Recall that the Ronkin function is linear on each component of the amoeba complement. We will call the corresponding flat pieces of the graph of the Ronkin function *facets*. They correspond to conical singularities (commonly referred to as "cusps") of the surface tension σ. The gradient of the Ronkin function maps $\mathbb{R}^2$ to the Newton polygon $N(P)$. It is known that the slopes of the facets form a subset of the integer points inside $N(P)$ [14]. Therefore, we have the following immediate corollary:

COROLLARY 3.7. *The surface tension σ is strictly convex and is smooth on the interior of $N(P)$, except at a subset of points in $\mathbb{Z}^2 \cap N(P)$. Also, σ is a piecewise linear function on $\partial N(P)$, with no slope discontinuities except at a subset of points in $\mathbb{Z}^2 \cap \partial N(P)$.*

COROLLARY 3.8. *The slope of $\tilde{\mu}(B_x, B_y)$ is the image of (B_x, B_y) under the map ∇F. That is, $\tilde{\mu}(B_x, B_y) = \mu(s,t)$ where $(s,t) = \nabla F(B_x, B_y)$.*

Proof. Since $\tilde{\mu}(B_x, B_y)$ is an EGM, it is equal to $\mu(s,t)$ for some (s,t) by Theorem 2.1. By (6) we must have $(s,t) = \nabla F(B_x, B_y)$. □

In Section 5, we will see that the spectral curves of dimer models are always very special real plane curves. As a result, their amoebas and Ronkin function have a number of additional nice properties, many of which admit a concrete probabilistic interpretation.

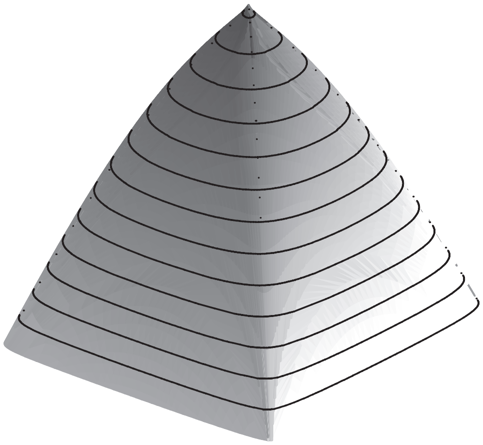

Figure 7. (*negative of*) *Surface tension for the square-octagon graph*

Figure 7 shows the Legendre dual of the Ronkin function from Figure 5. It is the surface tension function for certain periodically weighted dimers on the square grid with 2×2 fundamental domain and also for the uniformly weighted dimers on the square-octagon graph.

4. Phases of the dimer model

4.1. *Frozen*, *liquid*, *and gaseous phases.* We will show that EGMs with distinctly different qualitative properties are possible in a general periodic dimer model. The different types of behavior can be classified as frozen, liquid, and gaseous. We will take the fluctuation of the height function as the basis for the classification. This, as it will turn out, is equivalent to the classification by the rate of decay of correlations.

Let f and f' be two faces of the graph G and consider the height function difference $h(f) - h(f')$. An EGM is called a *frozen phase* if some of the height differences are deterministic—i.e., there exist distinct f and f' arbitrarily far apart for which $h(f) - h(f')$ is deterministic. An example is the delta-measure on the brick-wall matching of the square grid.

A nonfrozen EGM μ is called a *gaseous phase* or *smooth phase* if the height fluctuations have bounded variance, i.e., the μ variance of the random variable $h(f) - h(f')$ is bounded independently of f and f'. A nonfrozen EGM μ is called a *liquid phase* or *rough phase* if the μ-variance of the height difference is not bounded. The difference between the smooth and rough phases is illustrated in Figure 8.

We will prove in Theorem 4.5 that in the liquid phase the variance of $h(f) - h(f')$ grows *universally* like π^{-1} times the logarithm of the distance between f and f'. The following is our main result about phases:

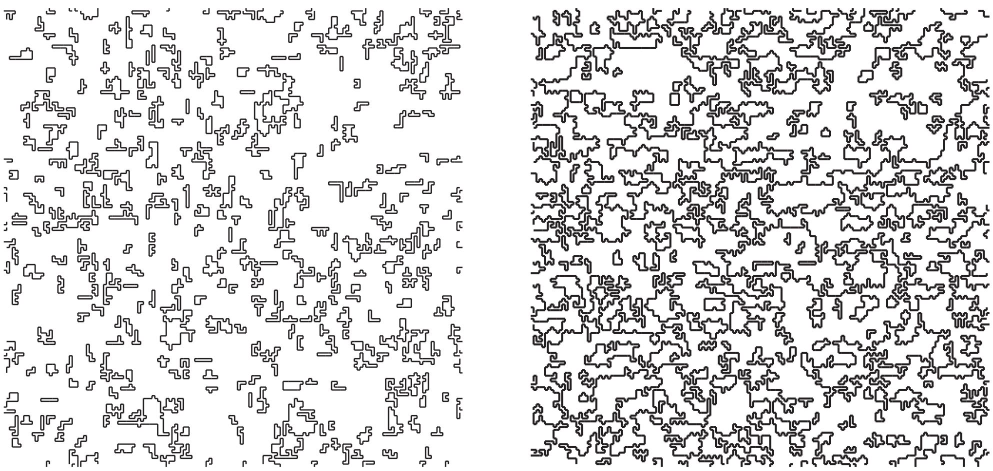

Figure 8. *All cycles of length ten or longer in the union of two random perfect matching of* $\mathbb{Z}^2$ *with* 4×4 *fundamental domain. The weight of one edge equals to* 10 *and all other edges have weight* 1. *The amoeba for this case is plotted in* Figure 6. *The slope on the left is* $(0, 0)$ *and this is a smooth phase. The rough phase on the right has slope* $(0, 0.5)$.

Theorem 4.1. *The measure* $\tilde{\mu}(B_x, B_y)$ *is frozen, liquid, or gaseous, respectively, when* (B_x, B_y) *is respectively in the closure of an unbounded complementary component of* $\mathbb{A}(P)$, *in the interior of* $\mathbb{A}(P)$, *or in the closure of a bounded complementary component of* $\mathbb{A}(P)$.

This theorem is proved in the next three sections. In Corollary 9.1.2 of [18] there is a different proof that when (s, t) lies in the interior of $N(P)$, $\mu(s, t)$ can only be smooth if s and t are integers.

We will see that in the liquid and gaseous phases the edge-edge correlations decay polynomially and exponentially, respectively. In the frozen case, some edge-edge correlations do not decay at all.

4.2. *Frozen phases.*

4.2.1. *Matchings and flows.* Recall the interpretation of a matching as a black-to-white unit flow. If M is a matching and M_0 is the reference matching in the definition of the height function, then the difference of the flows $M - M_0$ defines a divergence-free flow. The height function of M is the corresponding flux, that is, for two faces f_1, f_2, $h(f_2) - h(f_1)$ is the amount of flow crossing any dual path from f_1 to f_2. For two adjacent faces f_1, f_2 let $d(f_1, f_2)$ be the maximal possible (oriented) flow along the edge e between them (where e is oriented so that f_1 is on its left). This is the forward capacity of the oriented edge e. That is, if $e \notin M_0$, its capacity is 1 from its black vertex to its white vertex, and 0 in the other direction; if $e \in M_0$, its capacity is 1 from its white vertex to its black vertex, and zero in the reverse direction. For any two faces f_1 and f_2 let $D(f_1, f_2)$ be the minimum, over all dual paths from f_1 to f_2, of the sum of the capacities of the segments oriented to cross the path from left to right. By the max-flow-min-cut theorem, a function h is the height function for a tiling if and only if

$$\text{for all } \ f_1, f_2 \qquad D(f_1, f_2) \geq h(f_2) - h(f_1).$$

See [3], [21] for a reference.

Now let $(s,t) \in \mathbb{R}^2$. If there is no tiling with height function having slope (s,t) then there is a face f and $(x,y) \in \mathbb{Z}^2$ such that $D(f, f+(x,y)) < sx+ty$. We claim that in this case there is a face path from f to some translate $f+(x',y')$ on which $D(f, f+(x',y')) < sx'+ty'$ and all faces along this path are of distinct types, that is, are not translates of each other (except for the first and last faces). To see this, note that if a face path $f_1, f_2, \ldots, f_k$ passes through two faces of the same type, say f_i and f_j, then one of the two paths $f_1, \ldots, f_i, (f_i - f_j) + f_{j+1}, (f_i - f_j) + f_{j+2} \ldots, (f_i - f_j) + f_k$ and $f_i, \ldots, f_j$ will necessarily satisfy the strict inequality.

But up to translation there are only a finite number of face paths which start and end at the same face type and which pass through each face type at most once. Each such path gives one restriction on the slope: $D(f_1, f_2) \geq sx + ty$ where $(x,y) = f_2 - f_1$.

In particular the Newton polygon $N(P)$ is the set of (s,t) defined by the intersection of the inequalities $\{(s,t) \mid sx + ty \leq D(f_1, f_2)\}$, one for each of the above finite number of paths. If (s,t) is on the edge of $N(P)$, the path γ from the corresponding inequality has maximal flow, that is, in a tiling of slope (s,t) all edges on γ are determined: they must occur with probability 1 or 0.

4.2.2. *Frozen paths.* When (B_x, B_y) is in an unbounded component of the complement of the amoeba, we prove that $\tilde{\mu}(B_x, B_y)$ is in a frozen phase.

The slope (s,t) of $\tilde{\mu}(B_x, B_y)$ is an integer point on the boundary of $N(P)$. By the argument of the previous section, there is a face path γ on G_1, with homology class perpendicular to (s,t), for which every edge crossing each lift of γ is present with probability 1 or 0 for $\tilde{\mu}(B_x, B_y)$. These lifts constitute *frozen paths* in the dual G'. Edges which are in different components of the complement of the set of frozen paths are independent.

For each corner of $N(P)$ there are two sets of frozen paths, with different asymptotic directions. The components of the complements of these paths are finite sets of edges. The edges in each set are independent of all their translates.

4.3. *Edge-edge correlations.* For a finite planar graph Γ the inverse of the Kasteleyn matrix determines the edge probabilities: the probability of a set of edges $\{e_1,\dots,e_k\}$ being in a random matching is the determinant of the corresponding submatrix of K^{-1}, times the product of the edge weights [10].

For a graph on a torus the corresponding statement is more complicated: We have

THEOREM 4.2 ([2]). *The probability of edges* $\{e_1 = (\mathsf{w}_1,\mathsf{b}_1),\dots,e_k = (\mathsf{w}_k,\mathsf{b}_k)\}$ *occurring in a random matching of* G_n *equals* $\prod K(\mathsf{w}_j,\mathsf{b}_j)$ *times*

$$\frac{1}{2}\left(\frac{-Z_n^{(00)}}{Z}\det(K_{00}^{-1}(\mathsf{b}_j,\mathsf{w}_i)) + \frac{Z_n^{(10)}}{Z}\det(K_{10}^{-1}(\mathsf{b}_j,\mathsf{w}_i)) + \frac{Z_n^{(01)}}{Z}\det(K_{01}^{-1}(\mathsf{b}_j,\mathsf{w}_i)) + \frac{Z_n^{(11)}}{Z}\det(K_{11}^{-1}(\mathsf{b}_j,\mathsf{w}_i))\right). \tag{7}$$

Here the determinants $\det(K_{\theta\tau}^{-1}(\mathsf{b}_j,\mathsf{w}_i))$ are $k\times k$ minors of $K_{\theta\tau}^{-1}$. The asymptotics of this expression are again complicated by the zeros of P on $\mathbb{T}^2$. The entries in $K_{\theta\tau}^{-1}$ have the form (see [2])

$$K_{\theta\tau}^{-1}(\mathsf{b},\mathsf{w}) = \frac{1}{n^2}\sum_{z^n=(-1)^\theta}\ \sum_{w^n=(-1)^\tau}\frac{Q(z,w)\,w^x z^y}{P(z,w)}$$

where $Q(z,w)$ is one of a finite number of polynomials (depending on where w and b sit in their respective fundamental domains: Q/P is an entry of $K(z,w)^{-1}$ where $K(z,w)$ is the magnetically altered Kasteleyn matrix) and $(x,y)\in\mathbb{Z}^2$ is the translation taking the fundamental domain containing w to the fundamental domain containing b.

This expression is a Riemann sum for the integral

$$\frac{1}{(2\pi i)^2}\int_{\mathbb{T}^2}\frac{Q(z,w)w^x z^y}{P(z,w)}\frac{dw}{w}\frac{dz}{z},$$

except near the zeros of P. However the contribution for the root (z,w) nearest to a zero of P is negligible unless (z,w) is at distance $O(\frac{1}{n^2})$ of the zero. But if this is the case then replacing n with any n' at distance at least $O(\sqrt{n})$ from n makes the contribution for this root negligible. Thus we see that

$$\lim_{n\to\infty}{}' K_{n,\theta\tau}^{-1}(\mathsf{w},\mathsf{b}) = \frac{1}{(2\pi i)^2}\int_{\mathbb{T}^2}\frac{Q(z,w)w^x z^y}{P(z,w)}\frac{dw}{w}\frac{dz}{z},$$

where the limit is taken along a subsequence of n's.

Since all the $K_{n,\theta\tau}^{-1}$ have the same limit along a subsequence of ns, their weighted average (as in (7)) with weights $\pm Z_{\theta\tau}/2Z$ (weights which sum to one and are bounded between -1 and 1) has the same (subsequential) limit. This subsequential limit defines a Gibbs measure on $\mathcal{M}(G)$. By Theorem 2.1, this measure is the *unique* limit of the Boltzmann measures on G_n. Thus we have proved

Theorem 4.3. *For the limiting Gibbs measure $\mu = \lim_{n\to\infty}\mu_n$, the probability of edges $\{e_1,\dots,e_\ell\}$ where $e_j = (\mathsf{w}_j,\mathsf{b}_j)$, is*

$$\left(\prod_{j=1}^{\ell} K(\mathsf{w}_j,\mathsf{b}_j)\right)\det(K^{-1}(\mathsf{b}_k,\mathsf{w}_j))_{1\le j,k\le \ell}\,,$$

where, assuming b *and* w *are in a single fundamental domain,*

$$K^{-1}(\mathsf{b},\mathsf{w}+(x,y)) = \frac{1}{(2\pi i)^2}\int_{\mathbb{T}^2} K^{-1}(z,w)_{\mathsf{bw}}\, w^x z^y \frac{dw}{w}\frac{dz}{z}. \tag{8}$$

We reiterate that $K^{-1}(z,w)_{\mathsf{bw}} = Q_{\mathsf{bw}}(z,w)/P(z,w)$, where Q_{bw} is a polynomial in z and w.

4.4. *Liquid phases* (*rough nonfrozen phases*).

4.4.1. *Generic case.* When (B_x,B_y) is in the interior of the amoeba, Theorem 5.1, below, shows that $P(e^{B_x}z, e^{B_y}w)$ either has two simple zeros on the unit torus or a real node on the unit torus (a real node is a zero of P, $(z_0,w_0) = (\pm1,\pm1)$ where, locally, P looks like the product of two lines,

$$P(z,w) = (\alpha_1(z-z_0)+\beta_1(w-w_0))(\alpha_2(z-z_0)+\beta_2(w-w_0))+O(z-z_0,w-w_0)^3.$$

Generically P will not have real nodes). In the case of simple zeros (see Lemma 4.4 below), $K^{-1}(\mathsf{b},\mathsf{w})$ decays linearly but not faster, as $|w-b|\to\infty$. This implies that the edge covariances decay quadratically:

$$\begin{aligned}\mathrm{Cov}(e_1,e_2) &:= \Pr(e_1 \text{ and } e_2) - \Pr(e_1)\Pr(e_2)\\ &= -K(\mathsf{w}_1,\mathsf{b}_1)K(\mathsf{w}_2,\mathsf{b}_2)K^{-1}(\mathsf{b}_2,\mathsf{w}_1)K^{-1}(\mathsf{b}_1,\mathsf{w}_2)\,.\end{aligned}$$

In Section 4.4.2 we show that in the case of a real node we have similar behavior.

Lemma 4.4. *Suppose that* $|z_0| = |w_0| = 1$, $\mathrm{Im}(-\beta w_0/\alpha z_0) > 0$, $x,y\in\mathbb{Z}$, *and* $R(z,w)$ *is a smooth function on* $\mathbb{T}^2$ *with a only a single zero, at* (z_0,w_0), *and satisfying*

$$R(z,w) = \alpha(z-z_0)+\beta(w-w_0)+O(|z-z_0|^2+|w-w_0|^2).$$

Then we have the following asymptotic formula for the Fourier coefficients of R^{-1}:

$$\frac{1}{(2\pi i)^2}\int_{\mathbb{T}^2}\frac{w^x z^y}{R(z,w)}\frac{dz}{z}\frac{dw}{w} = \frac{-w_0^x z_0^y}{2\pi i(x\alpha z_0 - y\beta w_0)} + O\left(\frac{1}{x^2+y^2}\right).$$

If $\mathrm{Im}(-\beta w_0/\alpha z_0) < 0$ *then we get the same answer but with opposite sign.*

Note that if R has k simple zeros, $1/R$ can be written as a sum of k terms, each of which is of the above form.

Proof. Replacing z with $e^{ia}z_0$ and w with $e^{ib}w_0$ we have

$$\alpha(z-z_0)+\beta(w-w_0)+O(\dots)=\alpha z_0 ia+\beta w_0 ib+O(\dots).$$

By adding a smooth function (whose Fourier coefficients decay at least quadratically) to $1/R$ we can replace $1/R$ with

$$\frac{1}{\alpha z_0 ia+\beta w_0 ib}.$$

The integral is therefore

$$\frac{w_0^x z_0^y}{(2\pi)^2 i}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{e^{i(xb+ya)}}{\beta w_0 b+\alpha z_0 a}dadb+O(\dots)$$
$$=\frac{w_0^x z_0^y}{(2\pi)^2 i}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{e^{i(xb+ya)}}{\beta w_0 b+\alpha z_0 a}dadb+O(\dots).$$

We first integrate over the variable a: the integrand is a meromorphic function of a with a simple pole in the upper half-plane if $\mathrm{Im}(-\beta w_0 b/\alpha z_0)>0$. Change the path of integration to a path from $-N$ to N followed by the upper half of a semicircle centered at the origin of radius N. The residue theorem then yields

$$\frac{w_0^x z_0^y}{2\pi\alpha z_0}\int_0^{\infty}e^{i(x-y\beta w_0/\alpha z_0)b}db=\frac{w_0^x z_0^y}{2\pi\alpha z_0}\frac{-1}{i(x-y\beta w_0/\alpha z_0)}=\frac{-w_0^x z_0^y}{2\pi i(x\alpha z_0-y\beta w_0)}$$

which gives the result.

In case $\mathrm{Im}(-\beta w_0/\alpha z_0)<0$ the above integral would be from $-\infty$ to 0, resulting in the opposite sign. □

We now compute the variance in the height function.

THEOREM 4.5. *Suppose that the zeros of P on $\mathbb{T}^2$ are simple zeros at (z_0,w_0) and $(\bar z_0,\bar w_0)$. Let α,β be the derivatives of $P(z,w)$ with respect to z and w at (z_0,w_0). Then the height variance between two faces f_1 and f_2 is*

$$\mathrm{Var}[h(f_1)-h(f_2)]=\frac{1}{\pi}\log|\phi(f_1)-\phi(f_2)|+o(\log|\phi(f_1)-\phi(f_2)|),$$

where ϕ is the linear mapping $\phi(x+iy)=x\alpha z_0-y\beta w_0$.

Since ϕ is a nondegenerate linear mapping, the above expression for the variance is equivalent to $\frac{1}{\pi}\log|f_1-f_2|+o(\log|f_1-f_2|)$. However it appears that a slightly finer analysis would improve the little-o error in the statement to $o(1)$, so we chose to leave the expression in the given form.

Proof. Define $\tilde h=h-\mathbb{E}(h)$. Let f_1,f_2,f_3,f_4 be four faces, all of which are far apart from each other. We shall approximate $(\tilde h(f_1)-\tilde h(f_2))(\tilde h(f_3)-\tilde h(f_4))$. To simplify the computation we assume that f_1 and f_2 are translates of each other, as well as f_3 and f_4. Let $f_1 = g_1, g_2, \ldots, g_k = f_2$ be a path of translates of f_1 from f_1 to f_2, with $g_{p+1} - g_p$ being a single step in $\mathbb{Z}^2$. Similarly let $f_3 = g_1', g_2', \ldots, g_\ell' = f_4$ be a path from f_3 to f_4. We assume that these paths are far apart from each other.

Then

$$(9) \qquad (\tilde{h}(f_1) - \tilde{h}(f_2))(\tilde{h}(f_3) - \tilde{h}(f_4)) \\ = \sum_{p=1}^{k-1} \sum_{q=1}^{\ell-1} (\tilde{h}(g_{p+1}) - \tilde{h}(g_p))(\tilde{h}(g_{q+1}') - h(g_q')).$$

We consider one element of this sum at a time. There are three cases to consider: when $g_{p+1} - g_p$ and $g_{q+1}' - g_q'$ are both horizontal, both vertical, and one vertical, one horizontal.

Since P only has zeros at (z_0, w_0) and its conjugate, $Q(z,w)/P(z,w)$ can be written as sum of two terms $1/R(z,w)$ where R is as in Lemma 4.4. Therefore for x, y large we have

$$\frac{1}{(2\pi i)^2} \int_{\mathbb{T}^2} \frac{Q(z,w) w^x z^y}{P(z,w)} \frac{dz}{z} \frac{dw}{w} = -2\mathrm{Im}\left(\frac{w_0^x z_0^y Q(z_0, w_0)}{2\pi(x\alpha z_0 - y\beta w_0)} \right) + O(\frac{1}{x^2 + y^2}).$$

Recall that since $P(z,w) = \det K(z,w)$, the matrix $Q(z,w)$ satisfies

$$Q(z,w)K(z,w) = P(z,w) \cdot \mathrm{Id}.$$

Since (z_0, w_0) is a simple zero of P, $K(z_0, w_0)$ has co-rank 1. In particular $Q(z_0, w_0)$ must have rank 1. We write $Q(z_0, w_0) = UV^t$ where $V^t K(z_0, w_0) = 0 = K(z_0, w_0)U$.

Let $a_i = (\mathsf{w}_i, \mathsf{b}_i)$ be the edges crossing a "positive" face path γ from g_p to g_{p+1}, that is, a face path with the property that each edge crossed has its white vertex on the left. Similarly let $a_j' = (\mathsf{w}_j', \mathsf{b}_j')$ be the edges crossing a positive face path γ' from g_q' to g_{q+1}'. Then

$$\begin{aligned} &\mathbb{E}((\tilde{h}(g_{p+1}) - \tilde{h}(g_p))(\tilde{h}(g_{q+1}') - \tilde{h}(g_q'))) \\ &\quad = \mathbb{E}(\sum_{i,j} (a_i - \bar{a}_i)(a_j' - \bar{a}_j')) \\ &\quad = \sum_{i,j} \mathbb{E}(a_i a_j') - \mathbb{E}(a_i)\mathbb{E}(a_j') \\ &\quad = -\sum_{i,j} K(\mathsf{w}_i, \mathsf{b}_i) K(\mathsf{w}_j', \mathsf{b}_j') K^{-1}(b_j', \mathsf{w}_i) K^{-1}(\mathsf{b}_i, w_j'). \end{aligned}$$

Assuming these faces g_p, g'_q are far apart, this is equal to

$$-\frac{1}{4\pi^2}\sum_{i,j} K(\mathsf{w}_i,\mathsf{b}_i)K(\mathsf{w}_j,\mathsf{b}_j) \\ \times\left(\frac{w_0^x z_0^y U_i V'_j}{x\alpha z_0 - y\beta w_0} - \frac{\bar{w}_0^x \bar{z}_0^y \bar{U}_i \bar{V}'_j}{x\bar{\alpha}\bar{z}_0 - y\bar{\beta}\bar{w}_0} + O\left(\frac{1}{x^2+y^2}\right)\right) \\ \times\left(\frac{w_0^{-x} z_0^{-y} U'_j V_i}{x\alpha z_0 - y\beta w_0} - \frac{\bar{w}_0^{-x} \bar{z}_0^{-y} \bar{U}'_j \bar{V}_i}{x\bar{\alpha}\bar{z}_0 - y\bar{\beta}\bar{w}_0} + O\left(\frac{1}{x^2+y^2}\right)\right).$$

When we combine the cross terms we get an oscillating factor $w_0^{2x}z_0^{2y}$ or its conjugate, which causes the sum (9) of these terms when we sum over p and q to remain small. So the leading term for fixed p, q is

(10)
$$-\frac{2}{4\pi^2}\mathrm{Re}\left(\frac{1}{(x\alpha z_0 - y\beta w_0)^2}\sum_{i,j} K(\mathsf{w}_i,\mathsf{b}_i)K(\mathsf{w}_j,\mathsf{b}_j)U_iV_iU'_jV'_j\right) \\ = -\frac{1}{2\pi^2}\mathrm{Re}\left(\frac{1}{(x\alpha z_0 - y\beta w_0)^2}(\sum_i K(\mathsf{w}_i,\mathsf{b}_i)U_iV_i)(\sum_j K(\mathsf{w}_j,\mathsf{b}_j)U'_jV'_j)\right).$$

Now we claim that if $f_2 - f_1 = (1,0)$ then

$$\sum_{i\in\gamma} K(\mathsf{w}_i,\mathsf{b}_i)U_iV_i = z_0\frac{\partial P}{\partial z}(z_0,w_0) = z_0\alpha$$

and when $f_2 - f_1 = (0,1)$ then

$$\sum_{i\in\gamma'} K(\mathsf{w}_i,\mathsf{b}_i)U_iV_i = w_0\frac{\partial P}{\partial w}(z_0,w_0) = w_0\beta,$$

and similarly for $f_4 - f_3$. To see this, note first that the function $K(\mathsf{w}_i,\mathsf{b}_j)U_iV_j$ is a function on edges which is a closed 1-form, that is, a divergence-free flow. In particular the sum $\sum_i K(\mathsf{w}_i,\mathsf{b}_i)U_iV_i$ is independent of the choice of face path (in the same homology class). We can therefore assume that the face paths γ, γ' are equal to either γ_x or γ_y according to whether they are horizontal or vertical. Suppose for example $f_2 - f_1 = (1,0)$ and differentiate $Q(z,w)K(z,w) = P(z,w)\cdot\mathrm{Id}$ with respect to z and evaluate at (z_0,w_0): we get

$$Q_z(z_0,w_0)K(z_0,w_0) + Q(z_0,w_0)K_z(z_0,w_0) = P_z(z_0,w_0)\cdot\mathrm{Id}.$$

Applying U from the right to both sides, using $K(z_0,w_0)U = 0$ and $Q(z_0,w_0) = UV^t$, and then multiplying both sides by z_0, this becomes

$$UV^t z_0 K_z(z_0,w_0)U = z_0P_z(z_0,w_0)U.$$

However $z_0K_z(z_0,w_0) = \sum_{\gamma_x} K(\mathsf{w},\mathsf{b})\mathsf{w}\otimes\mathsf{b} = \sum_{i\in\gamma} K(\mathsf{w}_i,\mathsf{b}_i)\mathsf{w}_i\otimes\mathsf{b}_i$, so we get

$$U\sum K(\mathsf{w}_i,\mathsf{b}_i)V_iU_i = z_0P_z(z_0,w_0)U$$

and since $U\neq 0$ the claim follows. The same argument works for γ_y.

Recall that (x,y) was the translation between the fundamental domains containing g_p and g'_q. In the sum (9), let $(x_1,y_1)\in\mathbb{Z}^2$ be the position of the fundamental domain of g_p and (x_2,y_2) that of g'_q. Let $z_1 = x_1\alpha z_0 - y_1\beta w_0$ and $z_2 = x_2\alpha z_0 - y_2\beta w_0$. The sum (9) becomes (up to lower order terms)

$$-\frac{1}{2\pi^2}\mathrm{Re}\int_{\phi(f_1)}^{\phi(f_2)}\int_{\phi(f_3)}^{\phi(f_4)}\frac{dz_1dz_2}{(z_1-z_2)^2}$$

where ϕ is the linear map $(x,y)\mapsto x\beta w_0 - y\alpha z_0$. This integral evaluates to

$$\frac{1}{2\pi^2}\mathrm{Re}\log\left(\frac{(\phi(f_4)-\phi(f_1))(\phi(f_3)-\phi(f_2))}{(\phi(f_4)-\phi(f_2))(\phi(f_3)-\phi(f_1))}\right).$$

To approximate the height variance $\sigma(\tilde h(f_2)-\tilde h(f_1))$, let f_3 be close to f_1 and f_4 close to f_2 (but still far enough apart on the scale of the lattice so that the above approximations hold). Then as $|f_2-f_1|\to\infty$ while $|f_3-f_1|$ and $|f_4-f_2|$ are remaining bounded, the variance is

$$\frac{1}{\pi^2}\log|\phi(f_1)-\phi(f_2)| + o(\log|\phi(f_1)-\phi(f_2)|). \qquad \square$$

4.4.2. *Case of a real node.* In this section we show how to modify the above proof in case P has a real node. This happens when the two simple zeros $(z_0,w_0),(\bar z_0,\bar w_0)$ merge into a single zero at one of the four points $(\pm1,\pm1)$, and $P_z = P_w = 0$ there. In this case the slope of the corresponding EGM is integral but the amoeba does not have a complementary component, The component is reduced to a point (and is therefore not "complementary").

The canonical example of this behavior is the square grid with uniform weights and a 2×2 fundamental domain; in this case $P = 4+z+z^{-1}+w+w^{-1}$, and there is a real node at $(z,w)=(-1,-1)$.

Since $K^{-1}(\mathsf{b},\mathsf{w})$ is a continuous function of the edge weights, so is the variance of the height between two points. When $P(z,w)$ has a node, at say $(z,w)=(1,1)$, the polynomial $\tilde P(z,w) = P(e^{B_x}z, e^{B_y}w)$ has two simple zeros on $\mathbb{T}^2$ as long as B_x, B_y are sufficiently close to but not equal to 0. So the height variance of P is the limit of the height variances of $\tilde P$ as $B_x,B_y\to 0$.

In fact suppose without loss of generality that the node is at $(z,w)=(1,1)$ and P has the expansion

$$P(z,w) = a(z-1)^2 + b(z-1)(w-1) + c(w-1)^2 + \dots,$$

where $a, b, c \in \mathbb{R}$ and $\ldots$ denotes terms of order at least 3. Then near the node a point on P satisfies either $z - 1 = \lambda(w - 1) + O(w - 1)^2$ or $z - 1 = \bar{\lambda}(w - 1) + O(w - w_0)^2$ where $\lambda, \bar{\lambda}$, which are necessarily nonreal, are the roots of $a + bx + cx^2 = 0$.

The proof of Theorem 4.5 is valid for $\tilde{P}$ except for one assertion, where we ignored the cross terms in equation (10). Indeed, when z, w are each close to 1 the cross terms do not oscillate. However we will show that $\sum_i K(\mathsf{w}_i, \mathsf{b}_i)U_i\bar{V}_i = o(|\alpha| + |\beta|)$ as (z, w) tends to the node. Along with the complex conjugate equation this proves that the cross terms make no contribution.

The cross terms of (10) give

(11)
$$-\frac{1}{2\pi^2}\frac{1}{|x\alpha - y\beta|^2}\left((\sum_i K(\mathsf{w}_i, \mathsf{b}_i)U_i\bar{V}_i)(\sum_j K(\mathsf{w}_j, \mathsf{b}_j)\bar{U}_j'V_j') + (\text{conjugate})\right)$$

where recall that $\alpha = P_z, \beta = P_w$ are tending to 0 at the node. We first show that $\sum_i K(\mathsf{w}_i, \mathsf{b}_i)U_i\bar{V}_i = O(|\alpha| + |\beta|)$, and similarly for its complex conjugate. Recall the equation $QK = P \cdot \text{Id}$. Differentiating with respect to z we find

$$Q_zK + QK_z = P_z \cdot \text{Id}.$$

At a point (z_0, w_0) on P we have $Q = UV^t$ and at $(\bar{z}_0, \bar{w}_0)$ we have $Q = \bar{U}\bar{V}^t$. Applying $\bar{U}$ to the right and evaluating in the limit as (z_0, w_0) tends to the node, we see that

$$UV^tK_z\bar{U} = 0$$

so that $0 = V^tK_z\bar{U} = \sum_i K(\mathsf{w}_i, \mathsf{b}_i)U_i\bar{V}_i$ at the node. Since U, V can be chosen polynomial in z, w the quantity $\sum_i K(\mathsf{w}_i, \mathsf{b}_i)U_i\bar{V}_i$ necessarily vanishes to order at least one at the node (as do α and β).

Let c_1 be the limit at the node of $\frac{1}{\alpha}\sum_i K(\mathsf{w}_i, \mathsf{b}_i)U_i\bar{V}_i$ when $g_{p+1} - g_p = (1, 0)$ and c_2 the same limit when $g_{p+1} - g_p = (0, 1)$, so we may write

$$\lim \frac{1}{\alpha}\sum_i K(\mathsf{w}_i, \mathsf{b}_i)U_i\bar{V}_i = c_1dx + c_2dy$$

at the node. The cross terms are then

$$-\frac{1}{2\pi^2}\frac{1}{|x - y\beta/\alpha|^2}\Big((c_1dx_1 + c_2dy_1)(\bar{c}_1dx_2 + \bar{c}_2dy_2) \\ + (\bar{c}_1dx_1 + \bar{c}_2dy_1)(c_1dx_2 + c_2dy_2)\Big).$$

A short computation now shows that, since $\alpha/\beta \notin \mathbb{R}$, this is not a closed 1-form in (x_1, y_1) or (x_2, y_2) unless $c_1 = c_2 = 0$.

However since the height function differences $h(f_1) - h(f_2)$ do not depend on the path $g_1, \ldots, g_k$, the cross terms should necessarily be a closed 1-form. So $c_1 = c_2 = 0$. We have proved

THEOREM 4.6. *Suppose P has a real node $(z_0, w_0) = (\pm 1, \pm 1)$ on the unit torus, and*

$$P(z,w) = a(z-z_0)^2 + b(z-z_0)(w-w_0) + c(w-w_0)^2 + \dots.$$

Then the height variance between two faces f_1 and f_2 is

$$\mathrm{Var}[h(f_1) - h(f_2)] = \frac{1}{\pi}\log|\phi(f_1) - \phi(f_2)| + o(\log|\phi(f_1) - \phi(f_2)|),$$

where ϕ is the linear mapping $\phi(x+iy) = xz_0 - y\lambda w_0$, λ being the root of $a + b\lambda + c\lambda^2 = 0$.

4.5. *Gaseous phases* (*smooth nonfrozen phases*). When (B_x, B_y) is in a bounded complementary component, $P(e^{-B_x}z, e^{-B_y}w)$ has no zeros on the unit torus. As a consequence $K^{-1}(\mathsf{b},\mathsf{w})$ decays exponentially fast in $|\mathsf{b}-\mathsf{w}|$.

PROPOSITION 4.7. *The height variance $\sigma(h(f_1) - h(f_2))$ is bounded.*

Proof. The height difference $h(f_1) - h(f_2))$ can be measured along any path from f_1 to f_2. Suppose a dual path from f_1 to f_2 is chosen so that each edge crosses an edge of G with black vertex on its left. (This is clearly possible if f_1 and f_2 share a black or white vertex, by simply moving counterclockwise or clockwise around that vertex; since we can connect any f_1 and f_2 by a path in which consecutive faces share vertices, it is possible in general.) Then the height difference is a constant plus the sum of the indicator functions of the edges on the path: $h(f_1) - h(f_2) = \sum(a_i - \bar{a}_i)$. Consider two such paths γ_1, γ_2, consisting of edges a_i and b_j respectively, which are close only near their endpoints. The height variance is then the sum of the covariances of a_i and b_j:

$$\sigma(h(f_1) - h(f_2)) = \sum_{i,j} \Pr(a_i \text{ and } b_j) - \Pr(a_i)\Pr(b_j).$$

However these covariances are exponentially small except when both a_i and b_j are near the endpoints f_1 or f_2. In particular the above summation is approximately a geometric series, which has sum bounded independently of the distance between f_1 and f_2. □

4.6. *Loops surrounding the origin.* Smooth phases are characterized by their bounded height variance or exponential decay of correlation. Here is another characterization of smooth phases in terms of loops in the union of two perfect matchings sampled independently.

THEOREM 4.8. *A nonfrozen* EGM *μ is smooth if and only if when two perfect matchings M_1 and M_2 are chosen independently from μ, there are almost surely only finitely many cycles in $M_1 \cup M_2$ that surround the origin.*

Similarly, a nonfrozen EGM *is rough if and only if when two perfect matchings M_1 and M_2 are chosen independently from μ, there are almost surely infinitely many cycles in $M_1 \cup M_2$ that surround the origin.*

Proof. Since having infinitely many cycles surround the origin is a translation-invariant event, it is clear that if μ is an EGM, then there are either μ-almost-surely infinitely many cycles or μ-almost-surely finitely many cycles surrounding the origin. If the former is the case, it is easy to see that the variance is unbounded; to see this, simply use the fact that, conditioned on the positions of the cycles, the two *orientations* of a given cycle (i.e., which alternative set of edges in the cycle belongs to which of the M_i) are equally likely, and orientations of the cycles are independent of one another. (Note that the orientation of the cycle determines whether the height difference of the two height functions goes up or down when we cross that cycle.)

Now suppose that there are almost surely only finitely many cycles surrounding the origin. Lemma 8.4.3 of [18] further implies that if two perfect matchings M_1 and M_2 are sampled independently from $\mu(s,t)$, then the union $M_1 \cup M_2$ almost surely contains no infinite paths. It follows that the height difference between the two height functions is constant on the infinite cluster of faces that are not enclosed in any loops. The proof can now be completed using the fact that the height difference distribution is log concave, so that the expected number of cycles surrounding the origin is finite. Specifically, Lemma 8.3.4 of [18] implies that μ is smooth with respect to the (differently formulated but actually equivalent) definition given in Chapter 8 of [18], and Lemmas 8.1.1 and 8.1.2 imply that the height difference variances remain bounded in this case. □

5. Maximality of spectral curves

5.1. *Harnack curves.* The characteristic polynomial $P(z,w)$ has real coefficients and, hence, the spectral curve $P(z,w)=0$ is a real plane curve—the zero set of a real bivariate polynomial (in fact, it is more natural to consider the spectral curve as embedded in the toric surface corresponding to the Newton polygon of P). While all smooth complex curves of given genus are topologically the same, the number and the configuration of the ovals of a real plane curve can be very different. In particular, there is a distinguished class of real plane curves, known as *Harnack curves*, which have the maximal number of ovals (for given Newton polygon) in the, so to speak, best possible position. These curves are also known as maximal curves and also as simple Harnack curves. The precise topological definition of a Harnack curve can be found in [14]; here we will use the following alternative characterization of a Harnack curve obtained in [15]. Namely, a curve $P(z,w)$ is Harnack if and only if the map from the curve to its amoeba is 2-to-1 over the amoeba interior (except for a finite number of real nodes where it is 1-to-1). The main result of this section is the following

Theorem 5.1. *For any choice of nonnegative edge weights the spectral curve $P(z,w)=0$ is a Harnack curve.*

Harnack curves form a very special and much studied class of curves. Several characterizations and many beautiful properties of these curves can be found in [14], [15] (see also [12]). We will see that several of them have a direct probabilistic interpretation.

5.2. *Proof of maximality.* Maximality is an important property and several proofs of it are available. In many respects, it resembles the notion of total positivity [4], [7] and the proof given below exploits this analogy. Proofs of maximality based on different ideas can be found in [12], [13].

First observe that by the two-to-one property, being Harnack is a closed condition (since being not Harnack is clearly open), hence it is enough to prove that spectral curve is Harnack for a generic choice of weights. Furthermore, any periodic planar bipartite graph can be obtained, after using "vertex expansions and contractions", as a limit case of the periodically weighted hexagonal lattice when some of the edge weights are zero. (If a graph has a degree-2 vertex, removing this vertex and gluing its neighbors into a single vertex results in a graph with the same dimer coverings. This is a vertex contraction; a vertex expansion is the reverse of this process.) It is therefore, enough to consider the case of generic periodic weights on the hexagonal lattice with $n\times n$ fundamental domain.

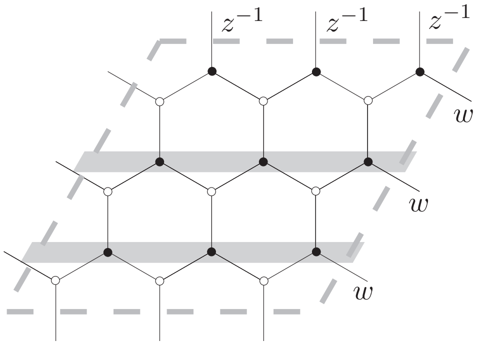

Figure 9. *The solution to $K(z,w)f=0$ can be constructed layer by layer*

By definition, $P(z,w)$ is a determinant of an $n^2\times n^2$ matrix $K(z,w)$ whose rows and columns are indexed by the n^2 white and black vertices in the fundamental domain. We can also write $P(z,w)$ as a determinant of an $n\times n$ matrix using transfer matrices as follows. The equation $\det K(z,w)=0$ means that there exists a nonzero function f on black vertices annihilated by the operator $K(z, w)$. We can construct such a function row by row as follows: given the values of f on a horizontal row of black vertices as in Figure 9, the equation $Kf = 0$ determines the values of f on the row below it. The corresponding linear map is given by $-T(w)$, where $T(w)$ is the transfer matrix of the form

$$T(w) = \begin{pmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & a_3 & b_3 & \\ & & & \ddots & \ddots \\ b_n\, w & & & & a_n \end{pmatrix}, \quad a_i, b_i > 0\,. \tag{12}$$

Here a_i are the weights on the edges in the NE-SW direction, b_i are weights on the NW-SE edges, and we may assume using a gauge transformation that the weights on the vertical edges are 1. Iterating this procedure once around the period, we get a consistency relation, which gives that $P(z, w) = 0$ only if

$$0 = P_2(z, w) = \det\left(z - (-1)^n\, T_1(w) \cdots T_n(w)\right)\,.$$

Since for generic weights both P and P_2 are both monic polynomials of degree d in z, with the same roots (and the roots are distinct, at least for w sufficiently large), we must have $P = P_2$.

Suppose now that for some point $(x, y) \in \mathbb{R}^2$ the torus $\{|z| = e^x, |w| = e^y\}$ contains more than two points of the curve $P(z, w)$. By changing the magnetic field, we can assume that $y = 0$. That is, we can assume that for a pair of points (z_1, w_1) and (z_2, w_2) on the spectral curve we have

$$|w_1| = |w_2| = 1\,, \quad w_2 \neq w_1, \bar{w}_1\,, \quad |z_1| = |z_2|\,.$$

Since being not Harnack is an open condition, we can find a nearby curve for which both w_1 and w_2 are roots of unity. It is easy to see (a more general statement is proved in [12]) that we can achieve this by a small perturbation of the dimer weights. So, we can assume that $w_1^m = w_2^m = 1$ for some integer m.

Taking nm as the new horizontal period, we find that the matrix

$$M = T_1(1) \cdots T_n(1) \tag{13}$$

has more than two eigenvalues of the same absolute value. We will now show that this is impossible.

This follows from the following lemma which is a version of a standard argument in the theory of total positivity (cf. [4]).

Lemma 5.2. *Suppose that all odd-size minors of a matrix M are nonnegative and that there exists k such that all odd-size minors of M^k are positive. Then the eigenvalues of M have the following form*

$$\lambda_1 > |\lambda_2| \geq |\lambda_3| > |\lambda_4| \geq |\lambda_5| > |\lambda_6| \geq |\lambda_7| > \ldots\,.$$

Remark, in particular, that if $|\lambda_{2k}| > |\lambda_{2k+1}|$ then both of these eigenvalues are real.

Proof. Let us order the eigenvalues of M so that

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \dots .$$

Since all matrix elements of M are nonnegative and all matrix elements of M^k are positive, the Perron-Frobenius theorem implies that λ_1 is simple, positive, and that

$$\lambda_1 > |\lambda_i| , \quad i > 1 .$$

Now consider the action of the matrix M in the third exterior power $\Lambda^3\mathbb{R}^n$ of the original space $\mathbb{R}^n$. The matrix elements of this action are the 3×3 minors of M and, hence, Perron-Frobenius theorem again applies. The eigenvalues of this action are the numbers

$$\lambda_i \lambda_j \lambda_k , \quad 1 \leq i < j < k \leq n .$$

It follows that the number $\lambda_1\lambda_2\lambda_3$ is real, positive, and greater in absolute value than $\lambda_1\lambda_2\lambda_i$ for any $i > 3$. It follows that

$$|\lambda_3| > |\lambda_4| .$$

Iteration of this argument concludes the proof. □

It is immediate to see that any matrix of the form $T_i(1)$ satisfies the hypothesis of the lemma. Matrices with nonnegative minors of any given order form a semigroup because, $s \times s$ minors of a matrix A are the matrix elements of A acting in the sth exterior power of the original space. It follows that all odd-size minors of (13) are nonnegative. In fact, all odd-size minors of M^k are positive for some large enough k, which can be seen as follows. Consider a graph with vertices indexed by s-element subsets of $\{1, \dots, n\}$. Two vertices are joined by an edge if the corresponding minor of M is positive. It is clear that this graph is connected and aperiodic, that is, some power of its adjacency matrix has positive entries. This concludes the proof.

5.3. *Implications of maximality.*

5.3.1. *Phase diagram of a dimer model.* The two-to-one property implies the following [14], [15]:

(i) the only singularities of the amoeba map (points where the map is not a local diffeomorphism) are folds over the boundary of the amoeba;

(ii) the boundary of the amoeba is the image of the real locus of the spectral curve;

(iii) to any lattice point in the interior of the Newton polygon corresponds either a bounded component of the amoeba complement or an isolated real node of the spectral curve. In particular, the number of holes in the amoeba equals the geometric genus of the curve[1].

For a general plane curve, more complicated singularities of the amoeba map are possible, which are then reflected in more complicated singularities of the Ronkin function and, hence, of its Legendre dual. In fact, we have used the two-to-one property in Section 4 in the classification of the phases of the dimer model. In this sense, the probabilistic meaning of maximality is the absence of any exotic phases with anomalous decay of correlations.

Part (iii) implies that the gaseous phases persist unless the corresponding component of the amoeba complement shrinks to a point and a nodal singularity develops. In particular, generically, the spectral curve is smooth and all gaseous phases are present. Conversely, if no gaseous phases are present then the spectral curve has the maximal possible number of nodes and hence is a curve of genus zero. It is shown in [12] that the latter case corresponds to isoradial dimers studied in [11]. It is also shown in [12] that all Harnack curves arise as spectral curves of some dimer model.

To summarize, part (iii) implies the following:

Theorem 5.3. *The number of gaseous phases of a dimer model equals the genus of the spectral curve. For a generic choice of weights, the dimer model has a gaseous phase for every lattice point in the interior of the Newton polygon $N(P)$ and a frozen phase for every lattice point on the boundary of $N(P)$.*

Part (ii) shows that the phase boundaries can be easily determined. Note that the outer oval (the component of the real locus intersecting the coordinate axes) of a Harnack curve is smooth and connected. It follows that any Harnack curve is irreducible, that is, the polynomial $P(z, w)$ is irreducible. Hence its amoeba can be defined by a single inequality

$$\prod P(\pm e^x, \pm e^y) \leq 0\,, \tag{14}$$

where the product is over all four choices of signs. Indeed, the product in (14) has a simple zero at the amoeba boundary (by irreducibility) and, hence, changes sign whenever we cross the boundary. Furthermore, generically, (14) is positive when x or y is large. We remark that for a general, non-Harnack, curve it is a rather nontrivial task to determine its amoeba, see the discussion in [20].

[1] For a nodal curve, the geometric genus equals $(2 - \chi - \#\text{ of nodes})/2$, where χ is the topological Euler characteristic.

Using the interpretation of $P(\pm 1, \pm 1)$ as the expectation of

$$(\pm 1)^{h_x}(\pm 1)^{h_y}(-1)^{h_x h_y}$$

with respect to the measure μ_1, see Section 3.1.3, we arrive at the following:

Theorem 5.4. *The minimal free energy measure μ is smooth* (*that is, frozen or gaseous*) *if the μ_1-measure of one of the four $H_1(\mathbb{T}^2, \mathbb{Z}/2\mathbb{Z})$ classes of matchings of G_1 exceeds $\frac{1}{2}$. If the μ_1-measure equals* $1/2$, *the measure is smooth unless the spectral curve has a real node over the origin in the amoeba* (*in which case μ is in a liquid phase*).

For example, for the square-octagon lattice, the spectral curve is $P = 5 + z + 1/z + w + 1/w$, see (1). The measure μ is smooth: the four homology classes have μ_1-weights proportional to $5 : 2 : 2 : 0$ (these are the weights of the coefficients of monomials of P with exponents taken modulo 2). For the square grid with 2×2 fundamental domain, $P(z, w) = 4 + z + 1/z + w + 1/w$. Even though the μ_1-weight of $0 \in H_1(\mathbb{T}^2, \mathbb{Z}/2\mathbb{Z})$ is $1/2$, the measure μ is in a liquid phase since $P(z, w) = 0$ has a real node at the origin.

5.3.2. *Universality of height fluctuations.* In Theorem 4.5 we proved that in a liquid phase the variance of the height function difference grows like π^{-1} times the logarithm of the distance. The proof of that theorem shows that the constant in front of the logarithm is directly connected to the number of roots of the characteristic polynomial on the unit torus. In particular, maximality was used in the essential way to show that this constant is always π^{-1}.

5.3.3. *Monge-Ampère equation for surface tension.* It follows from the results of [15] that the Ronkin function F of a Harnack curve satisfies the following Monge-Ampère equation

$$\det \begin{pmatrix} F_{xx} & F_{xy} \\ F_{yx} & F_{yy} \end{pmatrix} = \frac{1}{\pi^2}\,, \tag{15}$$

for any (x, y) in the interior of the amoeba. By the well-known duality for the Monge-Ampère equation, this implies the analogous equation for the surface tension function.

Theorem 5.5.

$$\det \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} = \pi^2\,. \tag{16}$$

It should be pointed out that in [6] it was argued that for certain class of random surface models the equation (16) should be satisfied at any *cusp* of the surface tension. It seems remarkable that in our case (16) is satisfied not just at a cusp but identically. For a general random surface model, we only expect the left-hand side of (16) to be positive, by strict concavity.

The geometric meaning of the equations (15) and (16) is that the gradients of F and σ, which are mutually inverse maps, are area-preserving, up to a factor. This leads to another characterization of Harnack curves as curves with amoebas of maximal possible area for given Newton polygon $N(P)$, namely π^2 times the area of $N(P)$. It would be interesting to find a probabilistic interpretation of this.

5.3.4. *Slopes and arguments.* Recall that by the Corollary 3.8 the gradient (s,t) of the Ronkin function at a point $(B_x, B_y) \in \mathbb{R}^2$ equals the slope of the measure $\tilde{\mu}(B_x, B_y)$. Suppose that the point (B_x, B_y) is in the interior of the amoeba and let (z_0, w_0) be one of its two preimages in the spectral curve. Maximality connects the slope (s,t) with the arguments z and w as follows.

THEOREM 5.6.

$$(s,t) = \pm\frac{1}{\pi}(\arg w_0, \arg z_0) \mod \mathbb{Z}^2 .$$

Proof.

$$s = \frac{d}{dB_x}\frac{1}{(2\pi i)^2}\iint_{\mathbb{T}^2} \log P(e^{B_x}z, e^{B_y}w)\frac{dz}{z}\frac{dw}{w}$$
$$= \frac{1}{2\pi i}\int_{|w|=e^{B_y}}\left(\frac{1}{2\pi i}\int_{|z|=e^{B_x}} d\log P\right)\frac{dw}{w} .$$

The inner integral counts, for w fixed, the number of zeros of $P(z,w)$ inside $\{|z| = e^{B_x}\}$. As w varies over the circle the number of zeros is locally constant with unit jumps whenever a zero crosses the circle. These jump points are precisely the points w where $P(z,w)$ has a root on $\mathbb{T}^2$, that is precisely the two points w_0 and $\bar{w}_0$ where (z_0, w_0) and its conjugate are the unique zeros of P on $\mathbb{T}^2$. Thus the integral is

$$\frac{1}{2\pi}\int_{-\arg w_0}^{\arg w_0} n d\phi + \int_{\arg w_0}^{2\pi-\arg w_0}(n\pm 1)d\phi = \pm\frac{\arg w_0}{\pi} \bmod \mathbb{Z}.$$

A similar argument applies for t. □

6. Random surfaces and crystal facets

The goal of this section is to review some known facts about random surfaces in order to say precisely what Theorem 3.6 and Theorem 4.1 imply about crystal facets and random surfaces based on perfect matchings. We begin with some analytical results.

6.1. *Continuous surface tension minimizers.* A standard problem of variational calculus is the following: given a bounded open domain $D \subset \mathbb{R}^2$ and any strictly convex surface tension function $\sigma : \mathbb{R}^2 \to \mathbb{R}$, find the continuous function $f : D \to \mathbb{R}$ whose (distributional) gradient minimizes the surface tension integral

$$I(f) = \int_D \sigma(\nabla f(x))\, dx$$

subject to the boundary condition that f extends continuously to a function f_0 on the boundary of D and the volume condition that

$$\int_D f(x) = B$$

for some constant B. If σ is the surface tension of one of the dimer models described in this paper, then $\sigma(u) = \infty$ whenever u lies outside of the closure of the Newton polygon $N(P)$, so that f is necessarily Lipschitz. The following result is well known (see [2], [18] for details and references).

Proposition 6.1. *If there exists any $\tilde{f}$, satisfying prescribed volume and boundary constraints and satisfying $I(\tilde{f}) < \infty$, then the surface tension minimizer f is unique and its gradient is almost everywhere defined.*

Both $\nabla\sigma$ and ∇f are functions from $\mathbb{R}^2$ to $\mathbb{R}^2$. The Euler-Lagrange equation for the functional $I(f)$ takes the following form:

Proposition 6.2. *Let f be the surface tension minimizer described above. Then whenever $x \in D$, f is C^2 at x, and σ is smooth at $f(x)$, we have*:

$$\mathrm{div}(\nabla\sigma \circ \nabla f(x)) = C \tag{17}$$

for some constant C, which depends on B and f_0. If f is also the minimal surface tension function when the volume constraint is ignored, then (17) *holds with $C = 0$.*

The archetypal solution to (17) is the Legendre dual of σ, which, in our setting, is the Ronkin function F. By construction

$$\nabla\sigma \circ \nabla F = \mathrm{Id}$$

throughout the amoeba—in particular, the divergence is constant. By analogy with the case of the Ronkin function we say that f has a *facet* of slope u if ∇f is equal to u on some open subset of D.

If f_0 is linear of any slope u on the boundary of D, then it is easy to see that the minimal surface tension function, ignoring the volume constraint, is linear of slope u, has (trivially) a facet of slope u, and satisfies (17) with $C = 0$. If we require, however, that $C \neq 0$, so that the volume constraint exerts some nonzero amount of pressure (upward or downward) on the surface, then Proposition 6.2 and Corollary 3.7 imply that the slope of any facet of f must be a lattice point inside $N(P)$. In other words, the facet slopes of the Ronkin function F represent all possible facet slopes of the dimer model.

6.2. *Concentration inequalities for discrete random surfaces.* In this section, we aim to show that the surface tension minimizing shapes and facets described in Section 6.1 are approximated by perfect-matching-based discrete random surfaces. First, suppose that D is a domain in $\mathbb{R}^2$, that f_0 is a continuous Lipschitz function defined on ∂D, and that f is a surface tension minimizer (with no volume constraint) which agrees with f_0 on the boundary.

Next, denote by $\frac{1}{n}G$ the weighted infinite graph G whose embedding into $\mathbb{R}^2$ has been re-scaled by a factor of $1/n$. For example, when $G = \mathbb{Z}^2$, then $\frac{1}{n}G$ is a grid mesh that is n times finer than G. Suppose that D_n is a sequence of simply connected subgraphs of $\frac{1}{n}G$ that *approximate* (D, f) *from the inside* in the sense that

1. The embedding of D_n is contained in D for all n.

2. The Hausdorff distance between the boundary of D_n and the boundary of D tends to zero in n.

3. Each D_n admits at least one perfect matching for which the height function h_n^0 on the boundary of D_n is such that $\sup |h_n^0 - f|$ tends to zero in n (where h_n^0 is treated as a function on a subset of the points in D, by letting $h_n(x)^0$ denote the height at the face of D_n containing x).

For each n, define ν_n to be the Boltzmann measure on perfect matchings M_n of D_n (i.e., the probability of each matching M is proportional to $\mathcal{E}(M)$, as defined in Section 2.2). Intuitively, we would expect that for sufficiently large values of n, the normalized function h_n/n (where h_n is sampled from ν_n) will closely approximate the continuous function f with high ν_n probability. A version of this statement is proved in [2] in the case $G = \mathbb{Z}^2$. Analogous but more general statements (in the form of *large deviations principles*) are discussed in Chapter 7 of [18] (see the paragraphs on Lipschitz potentials in Sections 7.3 through 7.5). In both [2] and [18], the results imply that, for a fixed value of ε, $\nu_n\{\sup|h_n/n - f| > \varepsilon\}$ tends to zero exponentially in n^2. Also, if $x \in D$ is a point at which $\nabla f(x) = u$, we conjecture that the local statistics of h_n, near the point x, are those of the Gibbs measure μ_{u_1,u_2}. More precise versions of this statement can be found in Chapter 7 [18]. The issue of weighting by enclosed volume is addressed briefly in Section 7.5 of [18].

University of British Columbia, Vancouver, B.C., Canada
E-mail address: kenyon@math.ubc.ca

Princeton University, Princeton, NJ
E-mail address: okounkov@math.princeton.edu

Courant Institute of Mathematical Science, New York University, NY
E-mail address: sheff@math.nyu.edu

References

[1] R. Cerf and R. Kenyon, The low-temperature expansion of the Wulff crystal in the three-dimensional Ising model, *Comm. Math. Phys.* **222** (2001), 147–179.

[2] H. Cohn, R. Kenyon, and J. Propp, A variational principle for domino tilings, *J. Amer. Math. Soc.* **14** (2001), 297–346 (electronic).

[3] J. C. Fournier, Pavage des figures planes sans trous par des dominos: fondement graphique de l'algorithme de Thurston et parallélisation, *C. R. Acad. Sci. Paris Sér.* I *Math.* **320** (1995), 107–112.

[4] F. Gantmacher and M. Krein, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, AMS Chelsea Publ., Providence, RI, 2002.

[5] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants*, *Resultants*, *and Multidimensional Determinants*, *Mathematics*: *Theory* & *Applications* (N. Wallach, ed.), Birkäuser Boston, Inc., 1994.

[6] P. Ferrari, M. Praehofer, and H. Spohn, Fluctuations of an atomic ledge bordering a crystalline facet; cond-mat/0303162.

[7] S. Karlin, *Total Positivity*, Stanford Univ. Press, Stanford, CA, 1968.

[8] P. Kasteleyn, Graph theory and crystal physics, in *Graph Theory and Theoretical Physics*, pp. 43–110, Academic Press, New York, 1967.

[9] R. Kenyon, An introduction to the dimer model, in *School and Workshop on Probability*, ICTP Lectures Notes (May, 2002) (G. Lawler, ed.), 2004; math.CO/0310326.

[10] ———, Local statistics of lattice dimers, *Ann. Inst. H. Poincaré*, *Probab. Statist.* **33** (1997), 591–618.

[11] ———, The Laplacian and $\bar{\partial}$ operators on critical planar graphs, *Invent. Math.* **150** (2002), 409–439.

[12] R. Kenyon and A. Okounkov, Planar Dimers and Harnack curves, *Duke Math. J.* **131** (2006), 499–524; math.AG/0311062.

[13] R. Kenyon and S. Sheffield, Dimers, tilings and trees, *J. Combin. Theory Ser.* B, to appear; math.CO/0310195.

[14] G. Mikhalkin, Amoebas of algebraic varieties; math.AG/0108225.

[15] G. Mikhalkin and H. Rullgård, Amoebas of maximal area, *Internat. Math. Res. Notices* **2001**, no. 9, 441–451.

[16] A. Okounkov, N. Reshetikhin, and C. Vafa, Quantum Calabi-Yau and classical crystals; hep-th/0309208.

[17] P. Pieranski, P. Sotta, D. Rohe, and M. Imperor-Clerc, Devil's staircase-type faceting of a cubic lyotropic liquid crystal, *Phys. Rev. Lett.* **84** (2000), 2409–2412.

[18] S. Sheffield, Ph.D. Thesis, Standford Univ., 2003, to appear, *Astérisque.*

[19] G. Tesler, *Matchings in graphs on non-orientable surfaces. J. Combin. Theory Ser.* B **78** (2000), 198–231.

[20] T. Theobald, Computing amoebas, *Experiment. Math.* **11** (2002), 513–526.

[21] W. P. Thurston, Conway's tiling groups, *Amer. Math. Monthly* **97** (1990), 757–773.

(Received February 6, 2004)