Chapter One

Measure, Energy, and Metric

1.1 GRAPH APPROXIMATIONS

In ordinary calculus we learn that continuous structures may be approximated by discrete structures. For example, the derivative is the limit of difference quotients, the integral is the limit of Riemann sums, and so on. At first, our naive intuition is that the discrete structures are simpler than the continuous ones, but we soon learn otherwise: The rules for derivatives are simpler than the corresponding rules for difference quotients (in fact, such rules as the product rule are rarely stated explicitly for difference quotients, although they do underlie the proofs of the corresponding derivative rules), and the fundamental theorem of the calculus allows very easy evaluation of some integrals. As we study calculus on fractals, we will also take the approach of using discrete approximations. At present there are no results that make the continuous structures simpler than the discrete ones, so we will have to devote careful attention to the discrete case. In the process we will learn some new things about ordinary calculus, since the unit interval is itself a self-similar set. Our plan is to develop the theory simultaneously for two examples: the unit interval I and the Sierpinski gasket SG.

The usual definition of the derivative involves arbitrary increments, and the Riemann sums in the definition of the integral allow arbitrary subdivisions of the interval. This is unnecessarily complicated. It suffices to deal with dyadic points $k/2^m$ $(0 \le k \le 2^m, 0 \le m < \infty)$. These points are dense in the interval, and as long as all the functions we deal with are continuous, it suffices to know the values at the dyadic points. To see how the dyadic points arise naturally we need to examine the self-similar structure of I. Consider the mappings $F_0 x = \frac{1}{2}x$ and $F_1 x = \frac{1}{2}x + \frac{1}{2}$ that send I to its left and right halves. Note that these are both contractive similarities (contraction ratio $\frac{1}{2}$, fixed points 0 and 1, respectively) and the images $F_0 I$ and $F_1 I$ intersect at the point $\frac{1}{2}$. The self-similar identity (the whole as union of similar parts)

$$I = F_0 I \cup F_1 I \tag{1.1.1}$$

uniquely determines I, as long as we specify that I is a nonempty compact set (both the empty set and the whole line satisfy (1.1.1), as well as the set of rational numbers in I, dyadic numbers in I, etc.). We note that (1.1.1) is not the only self-similar identity for I. For example, we can get many more by iteration. Write $F_w = F_{w_1} \circ F_{w_2} \circ \cdots \circ F_{w_m}$ for $w = (w_1, \ldots, w_m)$, each $w_j = 0$ or 1 (we call w a *word* of length $m = |w|$). Then

$$I = \bigcup_{|w|=m} F_w I \tag{1.1.2}$$

holds for any m. We will call this the *level m* decomposition and call $F_w I$ a *cell of level m*. Of course, (1.1.2) is just the decomposition of I into dyadic intervals $[k/2^m, (k+1)/2^m]$. We could also do irregular decompositions, such as

$$I = F_0 I \cup F_{10} I \cup F_{11} I. \tag{1.1.3}$$

There are also totally unrelated self-similar identities, for example involving $\frac{1}{3}x$, $\frac{1}{3}x + \frac{1}{3}$ and $\frac{1}{3}x + \frac{2}{3}$. This shows that the interval is different from the other fractals we will be studying.

The dyadic points are just the boundary points of the cells of various levels. Let us introduce some notation. $V_0 = \{q_0, q_1\}$ for $q_0 = 0$ and $q_1 = 1$ is the set of boundary points of I. Then inductively

$$V_m = \bigcup_i F_i V_{m-1}, \tag{1.1.4}$$

or equivalently

$$V_m = \bigcup_{|w|=m} \bigcup_i F_w q_i, \tag{1.1.5}$$

give the set of dyadic points $\{k/2^m\}$ for fixed m. Note that aside from the boundary points V_0, every point in V_m has two addresses, $x = F_w q_0 = F_{w'} q_1$, for the appropriate choices of w and w', so x is the left endpoint of one cell and the right endpoint of an adjacent cell. We will call such points *junction points*. We will regard the sets V_m as the vertices of a graph Γ_m, with edges written $x \underset{m}{\sim} y$ provided $x = k/2^m$ and $y = (k+1)/2^m$. Equivalently $x \underset{m}{\sim} y$ if there exists a cell of level m containing both x and y (as boundary points). Inductively, we build the graph Γ_m from the graph Γ_{m-1} by taking the two images $F_0\Gamma_{m-1}$ and $F_1\Gamma_{m-1}$ and identifying the common vertex $\frac{1}{2}$. See Figure1.1.1. Note that the set of vertices is increasing,

$$V_0 \subseteq V_1 \subseteq V_2 \subseteq \cdots. \tag{1.1.6}$$

However, the edge relations change: If x, y both belong to V_m and $x \underset{m}{\sim} y$, then x, y both belong to V_{m+1} but are *not* connected by an edge in Γ_{m+1}. Also note that every junction point in Γ_m has exactly two neighbors in V_m. Of course these graphs are very boring!

So now let's look at the case of SG. The self-similar structure of SG may be viewed as a natural generalization of the self-similar structure of I. This time we

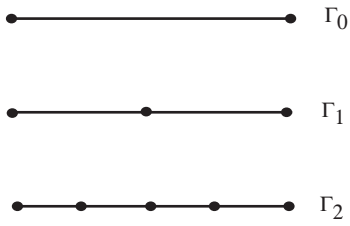

Figure 1.1.1

work in the plane and consider a set of three mappings $F_i : \mathbb{R}^2 \to \mathbb{R}^2$, $i = 0, 1, 2$, defined by

$$F_i x = \frac{1}{2}(x - q_i) + q_i, \tag{1.1.7}$$

where $\{q_i\}$ are the vertices of a triangle (any nondegenerate triangle will do). Then SG satisfies the self-similar identity

$$SG = \bigcup_{i=0}^{2} F_i(SG). \tag{1.1.8}$$

As in the case of I, the mapping F_i is a contractive similarity with contraction ratio $\frac{1}{2}$ and fixed point q_i. Also, SG is the unique nonempty compact set satisfying (1.1.8). The three cells on the right side of (1.1.8) intersect pairwise at single points. This means that while SG is connected, it is just barely so. If you remove just these three junction points, it becomes disconnected. You could think of SG as the ideal police state. To keep track of the whereabouts of all its citizens (at this level), the state need only post sentries at these three points. Similarly, if the state wants more detailed locations, it will post a finite number of sentries at more junction points. The terms "finitely ramified" and "postcritically finite" are used to describe this topological property. We will discuss the latter term in Chapter 4.

It is important to keep in mind that it is only the topological structure of SG that is of interest here, not the geometric structure inherited from its embedding in the plane. That is why we don't care which triangle we start with. But there are many other embeddings of SG in the plane, such as the famous Apollonian packing. We don't want to prejudice ourselves by looking at SG with "Euclidean eyes." In particular, although SG contains many straight line segments, we don't use any ordinary calculus concepts obtained by restricting functions on SG to these line segments. Eventually we will introduce a natural metric on SG that is not equivalent to the Euclidean metric (in any embedding in any dimensional Euclidean space) and that contains no rectifiable curves. Also, although our Euclidean eyes tend to see the triangle containing SG as a sort of boundary (since SG has no interior, the topological notion of boundary is not relevant), we will *define* the boundary to be the set $\{q_0, q_1, q_2\}$ of vertices of the triangle.

In order to be able to discuss I and SG simultaneously, we will use the symbol K to denote either one (and later other self-similar sets). The self-similar identities (1.1.1) and (1.1.8) may be combined as

$$K = \bigcup_i F_i K \tag{1.1.9}$$

(taking advantage of the ambiguity concerning the number of terms in the union). By iteration we have

$$K = \bigcup_{|w|=m} F_w K, \tag{1.1.10}$$

where F_w is defined as before, but the letters w_j in the word w may take on the values $\{0, 1, 2\}$ in the case of SG. This will be our decomposition of K into cells of

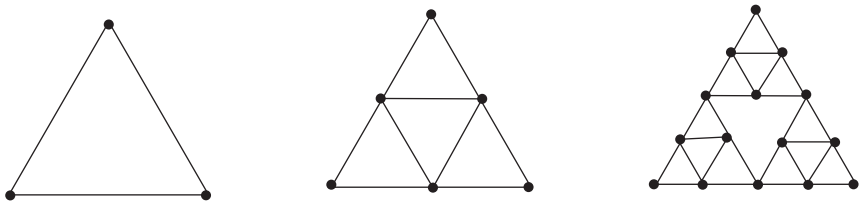

Figure 1.1.2

level m. Note that in the case of SG as well as I, distinct cells of level m are either disjoint or intersect at a single point; these will be our junction points. In the case of SG, unlike the interval, the junction points are topologically distinguishable from general points. In other words, while there are analogs of the decomposition (1.1.3), there are essentially no other decompositions. For SG we take $V_0 = \{q_0, q_1, q_2\}$. Then (1.1.4) or (1.1.5) defines V_m in both cases. Note that every point in $V_m \setminus V_0$ is a junction point with two addresses, $x = F_w q_i = F_{w'} q_{i'}$ for the appropriate choices (always $i \neq i'$). In particular,

$$\{F_0 q_1 = F_1 q_0,\ F_1 q_2 = F_2 q_1,\ F_2 q_0 = F_0 q_2\} \tag{1.1.11}$$

are the three junction points in $V_1 \setminus V_0$ in the SG case. We construct a graph Γ_m with vertices V_m by defining the edge relation $x \underset{m}{\sim} y$ if there is a cell of level m containing both x and y ($\exists\ w$ with $|w| = m$, i, j such that $x = F_w q_i$ and $y = F_w q_j$). In the case SG, Γ_m is obtained by taking the three copies $F_i \Gamma_{m-1}$ of Γ_{m-1} and identifying the points (1.1.11). Figure 1.1.2 shows the first three graphs.

Note that on SG every vertex in V_m, except for the three boundary points, has exactly four neighbors in Γ_m. There are times when the existence of the boundary is technically annoying, but we can easily get rid of it by passing to the double cover $\widetilde{K}$. That is, we take two copies of K and glue them together at the common boundary points. If we do this for I we obtain the circle, a one-dimensional manifold without boundary. The glued boundary points in $\widetilde{SG}$ have neighborhoods that are homeomorphic to neighborhoods of any junction point in SG. In this way, $\widetilde{SG}$ is an example of a "fractafold" without boundary modeled on SG. In the graphs $\widetilde{\Gamma}_m$ (two copies of Γ_m glued together at the corresponding boundary points), every vertex has exactly four neighbors. In other words, the graph is 4-regular.

EXERCISES

1.1.1. Show that $\#V_m = \frac{1}{2}(3^{m+1} + 3)$.

1.1.2. Let $x = F_w q_i$ with $|w| = m$ in $V_m \setminus V_0$. Give an algorithm for finding the other address $F_{w'} q_j$ for x in V_m. (Hint: If $w_m \neq i$, then $w'_k = w_k$ for $k < m$, $w'_m = i$, and $j = w_m$. If $w_m = i$, then $x \in V_{m-1}$, so reason by induction.)

1.1.3. Explicitly identify the four neighbors of x in Γ_m for $x \in V_m \setminus V_0$ and the two neighbors of q_i.

1.1.4. Show that the dihedral symmetry group D_3 of the equilateral triangle (three reflections, two rotations, and the identity) acts as a symmetry group of SG, and the action on Γ_m is given by permutations of the letters $\{0, 1, 2\}$.

1.1.5. Show that the set $V_1 \setminus V_0$ in SG is characterized topologically as the only set of three points whose removal disconnects SG into three components.

1.1.6. Show that SG is topologically rigid: Any homeomorphism must be one of the six symmetries in D_3.

1.1.7.* Show that $\widetilde{SG}$ is not topologically rigid in that there are infinitely many "accordian moves" across an identified boundary point.

1.1.8. Show that any two points in SG may be joined by a rectifiable curve (in fact, an infinite polygonal line).

1.1.9. (Nesting property) Show that if two cells (not necessarily of the same level) intersect in more than one point, then one contains the other.

1.1.10. If $x \in V_m$, then there exists a "chain" of points (not necessarily distinct) $x_0, x_1, \ldots, x_m$ such that $x_0 = q_0$, $x_k \in V_k$, and $x_m = x$, and $x_{k-1} \underset{k}{\sim} x_k$ for $1 \le k \le m$.

1.2 SELF-SIMILAR MEASURES

The notion of a general measure is a far-reaching generalization of notions such as length, area, volume, and probability. The theory is quite technical, and we will not attempt to describe it here. If you are familiar with measure theory, then you will be able to understand what we do here in that broader context (the key tool is the extension theorem of Carathéodory). If you are not familiar with measure theory you can still relax, because everything we are going to do is quite simple. This is thanks to the self-similar structure of K and also to the fact that we only need to integrate continuous functions, so we may imitate the integrals of Cauchy and Riemann rather than the integral of Lebesgue.

We want to consider what will be called here a *regular probability measure* μ on K. Roughly speaking, μ assigns weights $\mu(C)$ to all cells C of K in an additive fashion. Precisely, we require just the following four conditions:

(1.2.1) (positivity) $\mu(C) > 0$;

(1.2.2) (additivity) if $C = \bigcup_{j=1}^{N} C_j$,

where the cells $\{C_j\}$ intersect only at boundary points, then

$$\mu(C) = \sum_{j=1}^{N} \mu(C_j);$$

(1.2.3) (continuity) $\mu(C) \to 0$

as the size of C goes to 0;

(1.2.4) (probability) $\mu(SG) = 1$.

Condition (1.2.3) says that points have zero measure, and this enables us to ignore point intersections in condition (1.2.2). We may then extend the domain of μ to include finite unions of cells: If

$$A = \bigcup_{j=1}^{N} C_j \tag{1.2.5}$$

is a finite union of cells, disjoint except for point intersections, define

$$\mu(A) = \sum_{j=1}^{N} \mu(C_j). \tag{1.2.6}$$

Condition (1.2.2) guarantees that this is unambiguous: A different decomposition into cells would yield the same measure. To see this, first observe that there is a unique canonical decomposition of A into a union of cells of maximal size (C is not contained in a larger cell in A). Indeed, because of the nesting property, the maximal cells in A are disjoint (except for point intersections) and their union is A. Then any other representation of A is obtained by subdividing the maximal cells in some manner, and (1.2.2) says that the measure is conserved in the process. The additivity condition (1.2.2) continues to hold for sets that are finite unions of cells.

Similar reasoning shows that in place of (1.2.2) it suffices to verify for every cell $F_w K$,

$$\mu(F_w K) = \sum_i \mu(F_w F_i K), \tag{1.2.7}$$

the additivity for the decomposistion of $F_w K$ into cells of the next level. The construction of μ can then be imagined as follows: We assign weight 1 to SG, the cell of level 0. Inductively, having assigned weights to all cells of level m, we decide how to split the weight of each such cell when we subdivide it into cells at level $m+1$. The only restrictions are that (1.2.7) and (1.2.1) hold for the splitting, and (1.2.3) holds in the limit. Clearly there is a huge selection of measures!

The simplest choice is to do all splittings evenly. In the case of I, each cell of level m has measure 2^{-m}, its usual length. In fact, if A is any interval with dyadic endpoints, then $\mu(A)$ is the length of A. In the case of SG, each cell of level m will have measure 3^{-m}. We refer to this as the *standard measure*. It happens to coincide, up to a constant, with Hausdorff measure on SG in dimension $\log 3/\log 2$ (the exact value of the constant is an unsolved problem). Of course, the definition of Hausdorff measure is quite complicated, whereas our definition is quite simple.

The standard measure is a special case of a *self-similar* measure. To determine a self-similar measure we choose a set of probability weights $\{\mu_i\}$ on the index set $\{0, 1\}$ or $\{0, 1, 2, \}$,

$$\sum_i \mu_i = 1 \quad \text{with } \mu_i > 0, \tag{1.2.8}$$

and then set

$$\mu(F_w K) = \prod_{j=1}^{m} \mu_{w_j} \quad \text{for } |w| = m. \tag{1.2.9}$$

For simplicity we will write μ_w for the right side of (1.2.9). Another way of saying this is that we use the weights $\{\mu_i\}$ to accomplish the splitting (1.2.7). The standard measure is obtained by choosing all the μ_i equal, so $\mu_i = \frac{1}{2}$ for I and $\mu_i = \frac{1}{3}$ for SG.

For a self-similar measure, each mapping F_i contracts measures of sets by a factor μ_i,

$$\mu(F_i A) = \mu_i \mu(A), \tag{1.2.10}$$

since this is clearly true for cells by (1.2.9). Another way of expressing this is to take a set A and split it as $\bigcup_i A \cap F_i K$, and then by additivity,

$$\mu(A) = \sum_i \mu(A \cap F_i K). \tag{1.2.11}$$

But $F_i^{-1} A = F_i^{-1}(A \cap F_i K)$ and $A \cap F_i K = F_i F_i^{-1}(A \cap F_i K)$, so $\mu(A \cap F_i K) = \mu_i \mu(F_i^{-1}(A \cap F_i K))$ by (1.2.10). Together this shows $\mu(A \cap F_i K) = \mu_i \mu(F_i^{-1} A)$. Substituting into (1.2.11) yields the self-similar identity

$$\mu(A) = \sum_i \mu_i \mu(F_i^{-1} A). \tag{1.2.12}$$

It is easy to see that (1.2.12) implies (1.2.10) (replace A by $F_i A$, and then only one term survives on the right side) and hence (1.2.9), so the self-similar identity (1.2.12) (and the probability condition (1.2.4)) determines the measure μ uniquely. (It is also possible to prove this using a form of the contractive mapping principle, but the argument is more technical.)

On I, the self-similar measures are often called *Bernoulli measures*. Using the binary expansion, we may identify $x \in I$ with an infinite string of 0's and 1's (there is some ambiguity when x is a binary rational, but the set of binary rationals has measure zero and can be ignored). If we choose 0 with probability μ_0 and 1 with probability μ_1, independently for each binary digit, we get exactly the self-similar measure. Similarly, on SG we can interpret a self-similar measure as giving a recipe for choosing a point x in SG "at random." We first decide which of the three cells $F_i K$ of level 1 the point belongs to by spinning a roulette wheel where each outcome is alloted an angle of $2\pi\mu_i$. We call the chosen value w_1. We then determine which of the three level 2 cells $F_{w_1} F_i K$ x belongs to by another, independent spin of the same roulette wheel, and so on.

One of the main reasons for wanting to have a measure is to be able to integrate functions. Since the functions we want to integrate are usually continuous, this is easily accomplished by imitating the ordinary integral in calculus: We subdivide the space into a union of essentially disjoint small sets $\{A_j\}$ and take the limit of sums $\sum_j f(x_j)\mu(A_j)$, where $x_j \in A_j$. Since f is continuous, hence uniformly continuous on the compact space K, the choice of the point $x_j \in A_j$ does not matter in the limit.

In our setup there is a natural choice of subdivisions, namely

$$K = \bigcup_{|w|=m} F_w K, \tag{1.2.13}$$

the subdivision into all cells of level m. (See the exercises for other choices.) Then

$$\int_K f d\mu = \lim_{m\to\infty} \sum_{|w|=m} f(x_w)\mu(F_w K) \tag{1.2.14}$$

for $x_w \in F_w K$. It is not difficult to show that the limit exists and satisfies the usual properties of integrals: linearity in f, and the estimate

$$\min_K f \le \int_K f d\mu \le \max_K f. \tag{1.2.15}$$

If A is any finite union of cells, we can define $\int_A f d\mu$ by restricting to cells contained in A on the right side of (1.2.14). In analogy with the usual trapezoidal rule we may replace $f(x_w)$ by the average of f over the boundary points of the cell,

$$\int_K f d\mu = \lim_{m\to\infty} \frac{1}{3} \sum_{i=0}^{2} \sum_{|w|=m} f(F_w q_i)\mu(F_w K), \tag{1.2.16}$$

in the case of SG. This has the advantage of exhibiting the integral as a limit of discrete graph integrals. Given a graph, if we assign probabilities $\nu(x)$ to the vertices, we write

$$\int_\Gamma f d\nu = \sum_{x\in V} f(x)\nu(x). \tag{1.2.17}$$

Then (1.2.16) may be written

$$\int_K f d\mu = \lim_{m\to\infty} \int_{\Gamma_m} f d\nu_m, \tag{1.2.18}$$

where $\nu_m(x)$ is defined to be

$$\frac{1}{3}(\mu(F_w K) + \mu(F_{w'} K)) \text{ if } x \in V_m \setminus V_0$$

has the addresses $x = F_w q_i = F_{w'} q_j$, and $\frac{1}{3}\mu(F_i^m K)$ if $x = q_i$. For the standard measure this is simply

$$\int_{\Gamma_m} f d\nu_m = 3^{-m} \left(\frac{2}{3} \sum_{x\in V_m\setminus V_0} f(x) + \frac{1}{3} \sum_{x\in V_0} f(x) \right). \tag{1.2.19}$$

Note that we could drop the sum over the boundary points since this goes to zero in the limit. Nevertheless, for certain applications it is better to include them. The factor $\frac{2}{3}$ on the right side of (1.2.19) will play a role in the pointwise formula for the Laplacian.

In the case of the interval I, if we use the standard measure then we get the usual integral. For other choices of measure we get different integrals. There is a theorem that says $\int_I f d\mu = \int_0^1 f \circ \psi(x) dx$ for a suitable choice of a continuous change of variable function ψ (depending only on μ), but this function will be very irregular, certainly not differentiable.

If we fix a positive function f such that $\int_K f d\mu = 1$, then

$$\nu(A) = \int_A f d\mu \tag{1.2.20}$$

defines another measure. The measure ν is called *absolutely continuous* with respect to μ. In fact, the correct definition requires that we allow a much broader class of functions, including discontinuous and unbounded functions. You might wonder if it is possible to pass from one self-similar measure to another by such a construction, but in fact it is not possible. It is easy to see that it is impossible using a bounded function f, since there are many cells where the ratio $\mu'(C)/\mu(C)$ is larger than any fixed constant, if μ and μ' are distinct self-similar measures.

We observe that it is possible to transform the self-similar identity (1.2.12) into an identity involving integrals of functions. Indeed, if $f = \chi_A$, the characteristic function of the set A (not really continuous, but having only a finite set of discontinuities, so that the integral may be defined as before), then

$$\int_K f d\mu = \sum_i \mu_i \int_K f \circ F_i d\mu \tag{1.2.21}$$

is the same as (1.2.12) (note that $f \circ F_i = \chi_{F_i^{-1}A}$). By taking linear combinations and passing to the limit, it follows that (1.2.21) holds for all continuous functions.

EXERCISES

1.2.1. Show that the self-similar identity (1.2.12) generalizes to

$$\mu(A) = \sum_{|w|=m} \mu_w \mu(F_w^{-1} A) \quad \text{for any } m,$$

and similarly (1.2.21) generalizes to

$$\int_K f d\mu = \sum_{|w|=m} \mu_w \int_K f \circ F_w d\mu.$$

1.2.2. Let $\mathcal{P}$ be a finite set of words such that

$$K = \bigcup_{w \in \mathcal{P}} F_w K,$$

disjoint except for point intersections. We call $\mathcal{P}$ a *partition*. Show that

$$\mu(A) = \sum_{w \in \mathcal{P}} \mu_w \mu(F_w^{-1} A)$$

and

$$\int_K f d\mu = \sum_{w \in \mathcal{P}} \mu_w \int f \circ F_w d\mu.$$

1.2.3. Let $\rho = \min \mu_j$. Show that for any given $r > 0$ there exists a partition $\mathcal{P}$ such that $\rho r \le \mu_w \le r$ for every $w \in \mathcal{P}$.

1.2.4. Suppose $\mu_i \neq \mu_j$ for some $i \neq j$. Show that there exist adjacent cells $F_w K$ and $F_{w'} K$ with $|w| = |w'|$ such that $\mu(F_w K)/\mu(F_{w'} K)$ is as close to zero as desired.

1.2.5. Let μ be a self-similar measure on I. Use (1.2.21) to compute $\int_I x d\mu(x)$. Do the same for $\int_I x^2 d\mu(x)$.

1.3 GRAPH ENERGIES

Given a finite, connected graph G and a real-valued function u on its vertices, we define the graph energy by

$$E_G(u) = \sum_{x \sim y} (u(x) - u(y))^2. \tag{1.3.1}$$

Here the sum extends over all edges of the graph. If we were to sum first over all vertices x and then over all neighboring vertices, then each edge would occur twice and we would compensate by multiplying by a factor of $\frac{1}{2}$. Energy is a quadratic form in u. We will also need the associated bilinear form

$$E_G(u, v) = \sum_{x \sim y} (u(x) - u(y))(v(x) - v(y)) \tag{1.3.2}$$

for pairs of functions. Of course $E_G(u) = E_G(u, u)$, and we can recover the bilinear form from the quadratic form by the usual polarization identity:

$$E_G(u, v) = \frac{1}{4}(E_G(u + v) - E_G(u - v)). \tag{1.3.3}$$

It is clear that $E_G(u) = 0$ if u is constant, and the converse holds since we are assuming that G is connected. Also, $E_G(u, v)$ is an inner product on the space of functions on V modulo constants.

Another property of energy is called the *Markov property*: If we replace u by the minimum (or maximum) of u and a constant, the energy cannot increase. The reason for this is simply that each term $(u(x) - u(y))^2$ either stays the same or decreases. This is often stated in the form $E_G([u]) \leq E_G(u)$ for $[u] = \min\{1, \max\{u, 0\}\}$.

Now suppose we have two graphs, G and G', such that $V \subseteq V'$. We will think of G as a subgraph of G'. (We do not make any assumptions concerning the edges of G and G'.) Given a function u' on V', we can always restrict it to get a function $u = u'|_V$ on V. There is no apparent relationship between the energies $E_{G'}(u')$ and $E_G(u)$. If we go the other direction, starting with u defined on V, there are many possible extensions to V'. It is clear that there is at least one extension that minimizes the energy $E_{G'}(u')$. We will write $\tilde{u}$ for such an energy-minimizing extension and call it a *harmonic extension* (in the examples of interest, there will be a unique harmonic extension): $\tilde{u}|_V = u$ and $E_{G'}(\tilde{u}) \leq E_{G'}(u')$ for all u' such that $u'|_V = u$. We will call $E_{G'}(\tilde{u})$ the *restriction* of $E_{G'}$ to G.

Now it might happen, if we are lucky, that the restriction of $E_{G'}$ to G is equal to a multiple of E_G,

$$E_{G'}(\tilde{u}) = r E_G(u), \tag{1.3.4}$$

for all functions u on V. We call (1.3.4) a *renormalization* equation. Typically, $0 < r < 1$. This means that if we renormalize the definition of energy on G' by multiplying by $1/r$, then the restriction to G gives the same value, $r^{-1}E_{G'}(\tilde{u}) = E_G(u)$, and since $\tilde{u}$ is an energy minimizer, we have

$$r^{-1}E_{G'}(u') \geq E_G(u) \tag{1.3.5}$$

for every extension u' of u. In other words, energy increases with extension, except in the case of harmonic extension, when it remains unchanged.

This might seem like wishful thinking, but it actually works for the sequences of graphs Γ_m approximating K in both cases, I and SG! Let's look at I first. The first graph Γ_0 just consists of the vertices $\{0, 1\}$ connected by an edge, while Γ_1 consists of three vertices $\{0, \frac{1}{2}, 1\}$ connected sequentially. So the energies are given explicitly by (we change notation for simplicity)

$$E_0(u) = (u(1) - u(0))^2 \tag{1.3.6}$$

and

$$E_1(u') = \left(u'(1) - u'\left(\frac{1}{2}\right)\right)^2 + \left(u'\left(\frac{1}{2}\right) - u'(0)\right)^2. \tag{1.3.7}$$

If u' is an extension of u, then $u'(1) = u(1)$ and $u'(0) = u(0)$, so the only issue is, What is $u'(\frac{1}{2})$? To minimize $E_1(u')$ and so obtain the harmonic extension, it seems obvious that we should take

$$\tilde{u}\left(\frac{1}{2}\right) = \frac{1}{2}(u(1) + u(0)), \tag{1.3.8}$$

the linear extension (if we set $u'(\frac{1}{2}) = x$ and find the value where the x-derivative vanishes, we obtain (1.3.8)). A simple computation then reveals that

$$E_1(\tilde{u}) = \frac{1}{2}E_0(u), \tag{1.3.9}$$

a renormalization equation with $r = \frac{1}{2}$.

But now consider what happens when we go from Γ_m to Γ_{m+1}. The vertices V_{m+1} consist of all points of the form $\frac{k}{2^{m+1}}$, and among them, those with k even belong to V_m, while those with k odd are new. If u is defined on V_m, the question of harmonic extension is, What is $\tilde{u}\left(\frac{k}{2^{m+1}}\right)$ when k is odd? At first, minimizing energy may seem like a global problem, but in fact it is entirely local! Fix an odd value, say $k = 2j + 1$. Then $\tilde{u}\left(\frac{2j+1}{2^{m+1}}\right)$ only appears twice in $E_{m+1}(\tilde{u})$, specifically in the terms

$$\left(u\left(\frac{2j+2}{2^{m+1}}\right) - \tilde{u}\left(\frac{2j+1}{2^{m+1}}\right)\right)^2 + \left(\tilde{u}\left(\frac{2j+1}{2^{m+1}}\right) - u\left(\frac{2j}{2^{m+1}}\right)\right)^2. \tag{1.3.10}$$

This is the identical problem to the minimization of (1.3.7), and has the identical solution: Interpolate linearly,

$$\tilde{u}\left(\frac{2j+1}{2^{m+1}}\right) = \frac{1}{2}\left(u\left(\frac{2j+2}{2^{m+1}}\right) + u\left(\frac{2j}{2^{m+1}}\right)\right). \tag{1.3.11}$$

Then the same computation that yielded (1.3.9) shows that (1.3.10) is equal to $\frac{1}{2}\left(u\left(\frac{2j+2}{2^{m+1}}\right) - u\left(\frac{2j}{2^{m+1}}\right)\right)^2$, and summing over j we obtain

$$E_{m+1}(\tilde{u}) = \frac{1}{2}E_m(u), \tag{1.3.12}$$

a renormalization equation with the same constant $r = \frac{1}{2}$.

We define the renormalized graph energies by

$$\mathcal{E}_m(u) = r^{-m}E_m(u), \tag{1.3.13}$$

for $r = 1/2$. For any function, $\{\mathcal{E}_m(u)\}$ is a nondecreasing sequence. It is in fact constant when u is a linear function. A linear function (at least on the set V_* of dyadic rationals) is uniquely determined by its boundary values $u(0)$ and $u(1)$ by repeated use of the local extension algorithm (1.3.11). This may seem banal, because linear functions are such easily understood objects, but it will help us to understand the less trivial analog on SG.

The renormalized energy may be written explicitly as

$$2^m \sum_{k=1}^{2^m}\left(u\left(\frac{k}{2^m}\right) - u\left(\frac{k-1}{2^m}\right)\right)^2 = \sum_{k=1}^{2^m}\left(\frac{u\left(\frac{k}{2^m}\right) - u\left(\frac{k-1}{2^m}\right)}{\frac{1}{2^m}}\right)^2 \frac{1}{2^m}.$$

If u is continuously differentiable, the mean value theorem allows us to write this as

$$\sum_{k=1}^{2^m}(u'(x_k))^2\frac{1}{2^m}$$

for $\frac{k-1}{2^m} \le x_k \le \frac{k}{2^m}$, a Riemann sum for the integral

$$\int_0^1 u'(x)^2 dx. \tag{1.3.14}$$

So in that case $\mathcal{E}_m(u)$ converges to (1.3.14). We can also look at the renormalized bilinear form $\mathcal{E}_m(u,v) = r^{-m}E_m(u,v)$ and see that

$$\lim_{m\to\infty}\mathcal{E}_m(u,v) = \int_0^1 u'(x)v'(x)dx$$

if u and v are both continuously differentiable. We already know that if u is linear then $\mathcal{E}_m(u)$ is constant, but we may also assert that $\mathcal{E}_m(u,v)$ is constant for any function v. Indeed $u'(x)$ is then a constant, namely $u(1)-u(0)$, so

$$\int_0^1 u'(x)v'(x)dx = (u(1)-u(0))\int_0^1 v'(x)dx = E_0(u,v),$$

and by splitting the integral at the points $k/2^m$ we also obtain $\int_0^1 u'(x)v'(x)dx = \mathcal{E}_m(u,v)$. We may also observe directly that

$$\begin{aligned}\mathcal{E}_1(u,v) &= 2\left(u(1)-u\left(\frac{1}{2}\right)\right)\left(v(1)-v\left(\frac{1}{2}\right)\right)\\ &\quad +2\left(u\left(\frac{1}{2}\right)-u(0)\right)\left(v\left(\frac{1}{2}\right)-v(0)\right)\\ &= (u(1)-u(0))\left[\left(v(1)-v\left(\frac{1}{2}\right)\right)+\left(v\left(\frac{1}{2}\right)-v(0)\right)\right]\\ &= \mathcal{E}_0(u,v)\end{aligned} \tag{1.3.15}$$

since $u(1)-u(\frac{1}{2}) = u(\frac{1}{2})-u(0) = \frac{1}{2}(u(1)-u(0))$, etc.

Next we consider the case of SG. To keep the computation as simple as possible, we will exploit the symmetry. Suppose u is defined on V_0 by $u(q_0) = 1$ and $u(q_1) = u(q_2) = 0$, so $\mathcal{E}_0(u) = 2$, and we want to extend u to V_1 to minimize energy. By symmetry we will have $\tilde{u}(F_0q_1) = \tilde{u}(F_0q_2) = x$ at the junction points near q_0 and $\tilde{u}(F_1q_2) = y$ at the junction point opposite q_0 in V_1, where x and y are to be determined (see Figure 1.3.1). Then

$$E_1(\tilde{u}) = 2(x-1)^2 + 2x^2 + 2y^2 + 2(x-y)^2 \tag{1.3.16}$$

is to be minimized. By calculus we set the x and y derivatives equal to zero, to obtain the pair of linear equations

$$\begin{cases} 4x = 1 + x + y, \\ 4y = 2x. \end{cases} \tag{1.3.17}$$

Note that these equations express the mean value property that the function value at each of the junction points is the average of the function values of the four neighboring points in the graph. The solution $x = \frac{2}{5}$, $y = \frac{1}{5}$ is clear by inspection. By symmetric we would get the same answer if we put the value 1 at any of the boundary vertices. Also, since we are minimizing a quadratic function, the minimizing equations are linear. So if the initial values of u on V_0 are a, b, c, then the harmonic extension $\tilde{u}$ satisfies the following "$\frac{1}{5}-\frac{2}{5}$ rule":

$$u(z) = \frac{2}{5}a + \frac{2}{5}b + \frac{1}{5}c \tag{1.3.18}$$

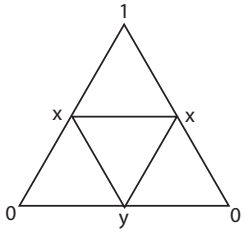

Figure 1.3.1 Values of $\tilde{u}$ on V_1.

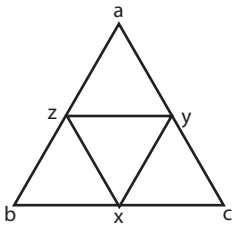

Figure 1.3.2 Values on V_1.

if z is the junction point between the vertices where u takes on the values a and b. Written this way, the harmonic extension will satisfy (1.3.18) on any cell of any level.

A more direct approach is to label the values on the vertices of V_1 as in Figure 1.3.2 and minimize E_1 as a function of x, y, z. The derivative equations yield the mean value equations

$$\begin{aligned} 4x &= b+c+y+z, \\ 4y &= a+c+x+z, \\ 4z &= a+b+x+y. \end{aligned}$$

Adding these equations yields

$$x+y+z = a+b+c,$$

so

$$5x = b+c+(x+y+z) = b+c+(a+b+c),$$

and so on.

Finally, we need to compute the renormalization factor. For the function in Figure 1.3.1 with $x = \frac{2}{5}$ and $y = \frac{1}{5}$, we find

$$\begin{aligned} E_1(\tilde{u}) &= 2\left(1-\frac{2}{5}\right)^2 + 2\left(\frac{2}{5}-\frac{1}{5}\right)^2 + 2\left(\frac{2}{5}-0\right)^2 + 2\left(\frac{1}{5}-0\right)^2 \\ &= \frac{18+2+8+2}{25} = \frac{6}{5}, \end{aligned}$$

so the choice $r = \frac{3}{5}$ yields

$$\mathcal{E}_1(\tilde{u}) = r^{-1}E_1(\tilde{u}) = E_0(u). \tag{1.3.19}$$

A little more work shows the same is true for the harmonic extension in the general case in Figure 1.3.2. Of course, a trivial remark is that the problem of minimizing the renormalized energy $\mathcal{E}_1$ is equivalent to minimizing E_1, with the same function $\tilde{u}$ achieving the minimum.

The same idea applies to the harmonic extension from V_1 to V_2, and in general from V_m to V_{m+1}. Suppose the values of u are given on V_m. Any new point in V_{m+1} (not in V_m) belongs to a unique m-cell F_wK with $|w| = m$. The total energy $E_{m+1}(u')$ for any extension is simply the sum of contributions from each cell F_wK,

$$E_{m+1}(u') = \sum_{|w|=m} (u'(F_wF_0q_0) - u'(F_wF_0q_1))^2 + \cdots = \sum_{|w|=m} E_1(u'\circ F_w),$$

and each contribution is just the energy E_1 of $u' \circ F_w$. So the global minimization problem is just the union of 3^m local minimization problems of the sort we have just solved. So the "$\frac{1}{5} - \frac{2}{5}$ rule" (1.3.18) continues to hold on each m-cell for the harmonic extension, and the renormalization factor is again $r = \frac{3}{5}$. Altogether, if we define

$$\mathcal{E}_m(u) = \left(\frac{3}{5}\right)^{-m} E_m(u), \tag{1.3.20}$$

then this renormalized energy remains unchanged under harmonic (minimum energy) extension, so it must go up for any extension:

$$\mathcal{E}_0(u) \le \mathcal{E}_1(u) \le \mathcal{E}_2(u) \cdots . \tag{1.3.21}$$

In the next section we will take the limit of this sequence.

To summarize what we have found so far: Given a function u on V_m, the harmonic extension $\tilde{u}$ to V_{m+1} may be characterized in three ways:

(i) it minimizes $\mathcal{E}_{m+1}(\tilde{u})$ at the value $\mathcal{E}_m(u)$;
(ii) at each new point $x \in V_{m+1} \setminus V_m$, $\tilde{u}(x)$ is the average of the values at the four neighboring points in V_{m+1};
(iii) it satisfies the "$\frac{1}{5} - \frac{2}{5}$ rule" at the new points in $V_{m+1} \setminus V_m$.

We may extend the equality in (i) to the bilinear form: If $\tilde{u}$, $\tilde{v}$ are the harmonic extensions of u and v, then

$$\mathcal{E}_{m+1}(\tilde{u}, \tilde{v}) = \mathcal{E}_m(u, v) \tag{1.3.22}$$

by the polarization identity (1.3.3), since harmonic extension is a linear transformation (from (iii)). As in the case of I, we can say more.

LEMMA 1.3.1 *Let u, v be defined on V_m, let $\tilde{u}$ be the harmonic extension of u, and let v' be any extension of v to V_{m+1}. Then*

$$\mathcal{E}_{m+1}(\tilde{u}, v') = \mathcal{E}_m(u, v). \tag{1.3.23}$$

Proof: Because of (1.3.22) it suffices to show $\mathcal{E}_{m+1}(\tilde{u}, v'') = 0$ for $v'' = v' - \tilde{v}$. Note that v'' vanishes on V_m. From the definition,

$$E_{m+1}(\tilde{u}, v'') = \sum_{x \underset{m+1}{\sim} y} (\tilde{u}(x) - \tilde{u}(y))(v''(x) - v''(y)).$$

Now collect all the terms that contain $v''(x)$ for a fixed x. If $x \in V_m$ then $v''(x) = 0$, so these terms contribute 0. But if $x \in V_{m+1}$, then $v''(x)$ multiplies

$$\sum_{y \underset{m+1}{\sim} x} (\tilde{u}(x) - u(y)), \tag{1.3.24}$$

and this vanishes by the mean value condition (ii). So $E_{m+1}(\tilde{u}, v'') = 0$ and hence $\mathcal{E}_{m+1}(\tilde{u}, v'') = 0$. □

Let's look at the "$\frac{1}{5} - \frac{2}{5}$" rule more closely. It says that the value at any inside point is a weighted average of the values of boundary points. The weight is higher

at the boundary points closest to the inside point, as is to be expected. I don't know of any explanatory argument for the exact values of the weights; they come from the computation. In Section 1.5 we will give another derivation for the value $r = \frac{3}{5}$, but it will also be the result of a different computation.

We define a *harmonic* function h to be one that minimizes $\mathcal{E}_m$ at all levels for the given boundary values on V_0. In other words, with $h(q_0), h(q_1), h(q_2)$ given, we inductively find $h\big|_{V_{m+1}}$ from $h\big|_{V_m}$ using the "$\frac{1}{5} - \frac{2}{5}$ rule." This is a local extension algorithm: If we want to zoom in to great depth in a small neighborhood, it is not necessary to compute h on the whole gasket. Specifically, if we want to know the values of h to level $m+k$ on the cell $F_w K$ for $|w| = m$, we only have to compute h on the cells $F_{w_1} K$, $F_{w_1 w_2} K$, $F_{w_1 w_2 w_3} K, \ldots, F_w K$ and then compute the values of h in complete detail for k more levels, for a total of $m + 3^k$ steps, as compared to 3^{m+k} steps for computing h on the whole gasket.

The space of harmonic functions, denoted $\mathcal{H}_0$, is three-dimensional. A simple basis $\{h_0, h_1, h_2\}$ is obtained by taking $h_j(q_j) = 1$ and $h_j(q_k) = 0$ for $k \neq j$. Certain properties of harmonic functions follow easily from the extension algorithm. Although h is initially defined only on V_*, it is uniformly continuous and so extends to a continuous function on K. It also satisfies the maximum principle: The maximum and minimum are attained on the boundary (and only on the boundary if the function is not constant). In the next chapter we will show that harmonic functions are exactly the solutions of the differential equation $\Delta h = 0$.

The renormalized energies $\mathcal{E}_m(h)$ are the same for all m, in particular for $m = 0$, so

$$\mathcal{E}_m(h) = (h(q_0) - h(q_1))^2 + (h(q_1) - h(q_2))^2 + (h(q_2) - h(q_0))^2. \tag{1.3.25}$$

In particular, $\mathcal{E}_m(h) > 0$ if h is nonconstant. Of course, if we start with h constant on V_0, then it remains constant on V_*, and it has zero energy by (1.3.25). In particular, $h_0 + h_1 + h_2 \equiv 1$.

It is convenient to represent the harmonic extension algorithm by a set of three matrices A_0, A_1, A_2 that describe how the boundary values change as we move from a cell of level m to its three subcells of level $m+1$. That is,

$$h|_{F_i V_0} = A_i h|_{V_0} \tag{1.3.26}$$

if we think of each set of h-values as a 3-vector, and more generally

$$h|_{F_w F_i V_0} = A_i h|_{F_w V_0}. \tag{1.3.27}$$

Indeed, (1.3.27) is just (1.3.26) applied to the function $h \circ F_w$, which is also a harmonic function. It is easy to see that

(1.3.28)
$$A_0 = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{5} & \frac{2}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \end{pmatrix}, \quad A_1 = \begin{pmatrix} \frac{2}{5} & \frac{2}{5} & \frac{1}{5} \\ 0 & 1 & 0 \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{pmatrix}, \quad A_2 = \begin{pmatrix} \frac{2}{5} & \frac{1}{5} & \frac{2}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \\ 0 & 0 & 1 \end{pmatrix}.$$

Another way of looking at it is that A_i is the matrix that represents the linear transformation $h \to h \circ F_i$ with respect to the basis $\{h_0, h_1, h_2\}$. Using the notation

$A_w = A_{w_m} \cdots A_{w_2} A_{w_1}$, we have

$$h|_{F_w V_0} = A_w h|_{V_0} \tag{1.3.29}$$

(if you are wondering about the correct order in the product, work out the case $m = 2$). It is important to understand that this is all there is! Unlike the case of the interval, there is no other description of harmonic functions. In principle it should be possible to obtain any desired information about harmonic functions from (1.3.29). In practice this may require a lot of work!

The individual matrices A_i are easy to understand. Each has eigenvalues 1, $\frac{3}{5}$, $\frac{1}{5}$. The eigenvector associated to 1 is the constant, but the eigenvectors associated to the other eigenvalues vary with the choice of i. For example, for A_0 the eigenvectors are $h_1 + h_2$ and $h_1 - h_2$ for eigenvalues $\frac{3}{5}$ and $\frac{1}{5}$, respectively. If we denote by R_0 the reflection symmetry that fixes q_0 and interchanges q_1 and q_2, then $h_1 + h_2$ is symmetric and $h_1 - h_2$ is skew-symmetric under R_0. If h is a harmonic function that vanishes at q_0, then it is a linear combination of $h_1 + h_2$ and $h_1 - h_2$ (write h as the sum of its symmetric and skew-symmetric parts). These functions have different decay rates as we approach q_0. Specifically, on the m-cell $F_0^m K$, $h_1 + h_2$ is $O((\frac{3}{5})^m)$ and $h_1 - h_2$ is $O((\frac{1}{5}))^m)$. A generic harmonic function vanishing at q_0 will have a nonzero symmetric part, so it will decay $O((\frac{3}{5})^m)$. To obtain the faster decay rate we have to choose a multiple of $h_1 - h_2$. In the next chapter we will see how to distinguish these cases by means of *normal derivatives*. The fact that the middle eigenvalue $\frac{3}{5}$ coincides with the renormalization constant r is no coincidence. The fact that $\frac{1}{5}$ is the smallest eigenvalue and 5 is the renormalization constant for the Laplacian is a coincidence. The numerology of these eigenvalues will have interesting consequences.

EXERCISES

1.3.1. The matrices A_i are invertible. Compute A_i^{-1} explicitly. Use these matrices to show how a harmonic function is uniquely determined by its values on the boundary of any given m-cell.

1.3.2. Consider the restriction of a harmonic function to the line segment in SG joining q_0 to q_1, and parametrize this segment by the unit interval in the obvious way. Find explicit formulas for $h(\frac{1}{4})$ and $h(\frac{3}{4})$ as a linear combination of $h(0)$, $h(\frac{1}{2})$, $h(1)$. Show that this algorithm localizes, so the values of h on all dyadic points in the interval (vertex points in the segment) are determined by $h(0)$, $h(\frac{1}{2})$, $h(1)$.

1.3.3.* Show that the restriction of h to this segment can have at most one local extremum.

1.3.4. Consider the two-dimensional space obtained from $\mathcal{H}$ by factoring out the constants. Choose a basis for this space and find explicit 2×2 matrices $\tilde{A}_i$ that represent the transformations $h \to h \circ F_i$ with respect to your basis. Note: There is no basis that is symmetric with respect to the dihedral group, so the result will not be as nice as (1.3.26), although each matrix $\tilde{A}_i$ will

have eigenvalues $\frac{3}{5}$, $\frac{1}{5}$ and can be made symmetric if the basis is chosen appropriately.

1.3.5. Note that for a nonconstant harmonic function,

$$\mathcal{E}_1(h) = r^{-1}\sum_{i=0}^{2}\mathcal{E}_0(h\circ F_i)$$

gives a decomposition of the energy of h into parts of the energy $r^{-1}\mathcal{E}_0(h\circ F_i)$ coming from each of the cells F_iK. Show that it is impossible to find h for which all these values $r^{-1}\mathcal{E}_0(h\circ F_i)$ are equal. More generally, describe all possible ways that the energy can be split.

1.3.6. Let $\operatorname{Osc}(f, A)$ denote the difference between the maximum and minimum values of f on A. Show that for harmonic functions $\operatorname{Osc}(h, F_iK) \le \frac{3}{5}\operatorname{Osc}(h, K)$, and more generally

$$\operatorname{Osc}(h, F_wK) \le \left(\frac{3}{5}\right)^m \operatorname{Osc}(h, K) \quad \text{if } |w| = m.$$

Use this to deduce that h on V_* is uniformly continuous.

1.3.7. Show that the eigenvalues of A_w for $|w| = m$ are $1, \lambda_1, \lambda_2$ where $\lambda_1\lambda_2 = (\frac{3}{25})^m$, and also

$$\left(\frac{1}{5}\right)^m \le |\lambda_1| \le |\lambda_2| \le \left(\frac{3}{5}\right)^m.$$

(Hint: Use the results of Exercise 1.3.4. The symmetry of $\tilde{A}_i$ implies the upper bound.)

1.3.8. Show that $\mathcal{E}_m(u^2) \le 4M^2\mathcal{E}_m(u)$ if $|u| \le M$. Prove a similar bound for $\mathcal{E}_m(uv)$ in terms of $\mathcal{E}_m(u)$, $\mathcal{E}_m(v)$ and upper bounds for u and v.

1.3.9.* Partition the edges of the graph Γ_m into three types, horizontal, slanting right, and slanting left, and similarly write $\mathcal{E}_m(h)$ as a sum of three "directional" energies $\mathcal{E}_m(h) = \mathcal{E}_m^{(1)}(h) + \mathcal{E}_m^{(2)}(h) + \mathcal{E}_m^{(3)}(h)$ by restricting the sum defining $\mathcal{E}_m$ to each type of edge. For harmonic functions h, show that $\mathcal{E}_m^{(1)}(h)$ converges to $\frac{1}{3}\mathcal{E}_0(h)$ as $m \to \infty$, for $i = 1, 2, 3$.

1.3.10. Show that

$$\begin{pmatrix} h_0\circ F_i \\ h_1\circ F_i \\ h_2\circ F_i \end{pmatrix} = A_i^* \begin{pmatrix} h_0 \\ h_1 \\ h_2 \end{pmatrix}.$$

1.4 ENERGY

In the previous section we constructed (for $K = I$ or SG) a sequence of energies $\mathcal{E}_m$ on Γ_m such that $\mathcal{E}_m(u)$ is increasing (nondecreasing) for any function u defined on V_*. It makes sense to define

$$\mathcal{E}(u) = \lim_{m\to\infty}\mathcal{E}_m(u), \tag{1.4.1}$$

allowing the value $+\infty$. Moreover, it is clear that $\mathcal{E}(u) = 0$ if and only if u is constant. We say $u \in \operatorname{dom}\mathcal{E}$ (u belongs to the domain of the energy) if and only

if $\mathcal{E}(u) < \infty$. We also say that u has *finite energy*. The definition of energy only involves the values of u on V_*, and we would really like to think of u as a function on K. We will see later that if u has finite energy then it is uniformly continuous on V_*, hence it has a unique continuous extension to K. By the way, this is not true in Euclidean spaces or manifolds of dimension 2 or more, so the graph approximation method does not work in those contexts.

In addition to showing that $\operatorname{dom}\mathcal{E} \subseteq C(K)$, we will show that $\operatorname{dom}\mathcal{E}$ is dense in $C(K)$, so that there exists an adequate supply of functions of finite energy. It is clear from the previous section that harmonic functions have finite energy, and an easy extension of this idea is that piecewise harmonic functions (start with any values on $V_{m'}$ for some fixed m' and extend harmonically for $m > m'$) also have finite energy. In fact, we will show that piecewise harmonic functions are dense, both in $C(K)$ and in $\operatorname{dom}\mathcal{E}$ in an appropriate sense.

Let u be a function of finite energy. Then $\mathcal{E}_m(u) \le \mathcal{E}(u)$, so if $x \underset{m}{\sim} y$ for $x, y \in V_m$ we have $r^{-m}(u(x)-u(y))^2 \le \mathcal{E}_m(u) \le \mathcal{E}(u)$ since $r^{-m}(u(x)-u(y))^2$ is a summand in $\mathcal{E}_m(u)$. This means

$$|u(x)-u(y)| \le r^{m/2}\mathcal{E}(u)^{1/2}. \tag{1.4.2}$$

This is already a statement of continuity. Now consider a chain of points x_m, $x_{m+1}, \dots, x_{m+k}$ such that $x_{m+j} \in V_{m+j}$ and $x_{m+j} \underset{m+j+1}{\sim} x_{m+j+1}$. Then we have

$$|u(x_m)-u(x_{m+k})| \le r^{m/2}(1+r^{1/2}+\cdots+r^{k/2})\mathcal{E}(u)^{1/2} \le \frac{r^{m/2}}{1-r^{1/2}}\mathcal{E}(u)^{1/2}$$

by adding up the estimates (1.4.2) along the chain of edges. From the geometry of K it is easy to see that if $x, y \in V_*$ belong to the same or adjacent m-cells, then we can connect x to y by at most two such chains, so

$$|u(x)-u(y)| \le \frac{2r^{m/2}}{1-r^{1/2}}\mathcal{E}(u)^{1/2}. \tag{1.4.3}$$

Not only is (1.4.3) a statement of uniform continuity, it is also a Hölder condition. In the case of the interval, if $|x-y| \le \frac{1}{2^m}$, then x and y belong to the same or adjacent m-cell. Since $r = \frac{1}{2}$, (1.4.3) says

$$|u(x)-u(y)| \le M|x-y|^{1/2}. \tag{1.4.4}$$

(This is the optimal Hölder condition in the Sobolev embedding theorem for H^1, which may be identified with $\operatorname{dom}\mathcal{E}$; see the exercises.) In the case of SG we also get a Hölder condition for the Euclidean metric with a strange exponent, $\log(\frac{5}{3})/\log 2$. In Section 1.6 we will introduce a more natural metric on SG, and with respect to this metric the Hölder exponent will again be $\frac{1}{2}$.

For the rest of this section, all functions will be assumed to be continuous and defined on all of K.

Lemma 1.4.1 *Let $u, v \in \operatorname{dom}\mathcal{E}$. Then*

$$\lim_{m\to\infty} \mathcal{E}_m(u,v) = \mathcal{E}(u,v) \tag{1.4.5}$$

*exists and defines an inner product on $\operatorname{dom}\mathcal{E}/$*constants.

Proof: We begin with the polarization identity

$$\mathcal{E}_m(u,v) = \frac{1}{4}(\mathcal{E}_m(u+v) - \mathcal{E}_m(u+v)) \tag{1.4.6}$$

at level m. Since the right side of (1.4.6) has a limit, so does the left side. The usual properties of an inner product, except that $\mathcal{E}(u) = 0$ may occur, follow easily. Since $\mathcal{E}(u) = 0$ implies that $\mathcal{E}_m(u) = 0$, which implies that u is constant on V_m for all m, it follows that u must be constant. By factoring out by the constants, we obtain a true inner product. □

THEOREM 1.4.2 *dom $\mathcal{E}$/constants forms a Hilbert space with inner product (1.4.5).*

Proof: It remains to show completeness: Every Cauchy sequence converges. It is convenient to identify dom $\mathcal{E}$/ constants with the space $\tilde{\mathcal{E}} = \{u \in \mathrm{dom}\,\mathcal{E} : u(q_0) = 0\}$. Let $\{u_n\}$ be a sequence in $\tilde{\mathcal{E}}$ such that $\mathcal{E}(u_n - u_{n'}) \to 0$ as $n, n' \to \infty$. Then for fixed m, $\mathcal{E}_m(u_n - u_{n'}) \to 0$ also, since $\mathcal{E}_m(u_n - u_{n'}) \le \mathcal{E}(u_n - u_{n'})$. It follows easily that

$$\lim_{n\to\infty} u_n(x) \quad \text{exists for each } x \in V_m,$$

so we may define u on V_* as this limit, and moreover

$$\mathcal{E}_m(u_n - u) = \lim_{n'\to\infty} \mathcal{E}_m(u_n - u_{n'}). \tag{1.4.7}$$

By taking n large enough, the right side of (1.4.7) may be made as small as desired independent of m, so $\mathcal{E}(u_n - u) \to 0$ as $n \to \infty$. □

Having to factor out by the constants is a minor nuisance. We will say $u_n \to u$ in energy if $\mathcal{E}(u_n - u) \to 0$ and also $u_n \to u$ uniformly (it suffices to have $u_n \to u$ at a single point in view of (1.4.2)).

DEFINITION 1.4.3 The space $S(\mathcal{H}_0, V_m)$ of piecewise harmonic splines of level m is defined to be the space of continuous functions such that $u \circ F_w$ is harmonic for all $|w| = m$.

It is easy to see that $S(\mathcal{H}_0, V_m)$ is contained in dom $\mathcal{E}$ and is a finite-dimensional space of dimension $\#V_m$. All such functions are obtained by specifying the values of u on V_m arbitrarily and then extending harmonically to $V_{m'}$ for each $m' > m$. Clearly $\mathcal{E}(u) = \mathcal{E}_m(u)$ for these functions.

THEOREM 1.4.4 *Any function $u \in C(K)$ may be approximated uniformly by a sequence $u_m \in S(\mathcal{H}_0, V_m)$, with $u_m\big|_{V_m} = u\big|_{V_m}$. Moreover, if $u \in$ dom $\mathcal{E}$ then u_m converges to u in energy.*

Proof: Given $\varepsilon > 0$, we can find m such that $\mathrm{Osc}(u, F_w K) \le \varepsilon$ for all w with $|w| = m$. Then since $u_m\big|_{V_m} = u\big|_{V_m}$ we also have $\mathrm{Osc}(u_m, F_w K) \le \varepsilon$, so

$$\begin{aligned} |u_m(x) - u(x)| &\le |u_m(x) - u_m(F_w q_0)| + |u_m(F_w q_0) - u(F_w q_0)| \\ &\quad + |u(F_w(q_0) - u(x)| \\ &\le 2\varepsilon \text{ for } x \in F_w K, \end{aligned}$$

so $\|u_m - u\|_\infty \le 2\varepsilon$.

Next suppose $u \in \text{dom } \mathcal{E}$. Then $\mathcal{E}_m(u) = \mathcal{E}_m(u_m) \nearrow \mathcal{E}(u)$. Also $\mathcal{E}(u, u_m) = \mathcal{E}_m(u, u_m) = \mathcal{E}_m(u_m)$ by Lemma 1.3.1. So

$$\mathcal{E}(u - u_m) = \mathcal{E}(u) - 2\mathcal{E}(u, u_m) + \mathcal{E}(u_m) = \mathcal{E}(u) - \mathcal{E}_m(u_m) \to 0.$$

□

The next result expresses the self-similarity of the energy.

THEOREM 1.4.5 *If $u \in dom\,\mathcal{E}$ then $u \circ F_i \in dom\,\mathcal{E}$ for all i, and*

$$\mathcal{E}(u) = \sum_i r^{-1}\mathcal{E}(u \circ F_i). \tag{1.4.8}$$

Proof: It is clear from the definition that

$$\mathcal{E}_{m+1}(u) = \sum_i r^{-1}\mathcal{E}_m(u \circ F_i).$$

Taking the limit we obtain (1.4.8), and if the left side is finite then each term on the right must also be finite. □

Of course the same identity holds for the bilinear form $\mathcal{E}(u, v)$ and for subdivisions

$$K = \bigcup_{w \in \mathcal{P}} F_w K \tag{1.4.9}$$

with r^{-1} replaced by $r^{-|w|}$, for any partition $\mathcal{P}$:

$$\mathcal{E}(u) = \sum_{w \in \mathcal{P}} r^{-|w|}\mathcal{E}(u \circ F_w). \tag{1.4.10}$$

Another way of saying this is that we can create a function of finite energy on K by gluing together finite energy functions on cells $F_w K$ provided the functions match at junction points.

The additivity in (1.4.10) suggests that we could think of energy as a measure. More precisely, define a measure ν_u by

$$\nu_u(F_w K) = r^{-|w|}\mathcal{E}(u \circ F_w). \tag{1.4.11}$$

Equivalently, $\nu_u(F_w K)$ is obtained as a limit of

$$\sum_{x \underset{m}{\sim} y} r^{-m}(u(x) - u(y))^2, \tag{1.4.12}$$

where the sum is restricted to those edges lying in $F_w K$. It is easy to check that all the conditions for a regular measure are satisfied except strict positivity (for a probability measure we would need $\mathcal{E}(u) = 1$). Then

$$\mathcal{E}(u) = \nu_u(K) = \int_K 1 d\nu_u. \tag{1.4.13}$$

An interesting difference beween I and SG is that on I, the energy measures are absolutely continuous with respect to the standard measure, but on SG they are not.

In fact they are singular—roughly speaking, they concentrate mass too much in neighborhoods of junction points. This result was first proved by Kusuoka. A hint of why this is so is given in Exercise 1.3.5.

Here are a couple of simple properties of energy: The Markov property

$$\mathcal{E}([u]) \le \mathcal{E}(u) \text{ for } [u] = \min\{1, \max\{u, 0\}\} \tag{1.4.14}$$

follows from the corresponding property for $\mathcal{E}_m$. Also, dom $\mathcal{E}$ forms an algebra under pointwise multiplication. We leave the verification to the exercises.

EXERCISES

1.4.1. Show that $\nu_u(F_w K) \le r^{|w|}\mathcal{E}(u)$ and so the continuity condition (1.2.3) holds for ν_u.

1.4.2. Show that $\operatorname{dom}\mathcal{E}$ is an algebra, and find an estimate for $\mathcal{E}(uv)$ in terms of $\mathcal{E}(u)$, $\mathcal{E}(v)$, $\|u\|_\infty$, and $\|v\|_\infty$.

1.4.3. Let R be one of the reflection symmetries in D_3, and suppose $u, v \in \operatorname{dom}\mathcal{E}$ are, respectively, symmetric and skew-symmetric with respect to R. Show that $\mathcal{E}(u, v) = 0$.

1.4.4. On SG choose an orthornormal basis $\{h_1, h_2\}$ for $\mathcal{H}_0$/constants with respect to the energy inner product, and define the *Kusuoka measure* $\nu = \nu_{h_1} + \nu_{h_2}$. Show that this measure is independent of the choice of orthonormal basis.

1.4.5. Show that if $u \in C^1(I)$ then $u \in \operatorname{dom}\mathcal{E}$ and $\mathcal{E}(u) = \int_0^1 (u'(x))^2 dx$.

1.4.6.* On I, show that $\operatorname{dom}\mathcal{E}$ may be identified with the Sobolev space H^1, with $\mathcal{E}(u) = \int_0^1 (u'(x))^2 dx$, where now u' is the distributional derivative and the integral is a Lebesgue integral.

1.4.7. Consider the skew-symmetric function u on SG defined by $u(F_0^k q_1) = 3^k$, $u(F_0^k q_2) = -3^k$, and extended to be harmonic on every cell not containing q_0 (see Figure 1.4.1). Show that u has infinite energy, but u is harmonic in the complement of q_0 (u satisfies the mean value condition at level m for any vertex in $V_m \setminus V_0$ not adjacent to q_0).

1.4.8. Show that if $u \in S(\mathcal{H}_0, V_m)$ and $v \in \operatorname{dom}\mathcal{E}$ then $\mathcal{E}(u, v) = \mathcal{E}_m(u, v)$.

1.4.9. (a) Show that the energy is *local*, meaning $u \cdot v \equiv 0$ implies $\mathcal{E}(u, v) = 0$. (b) Show that the energy is *strongly local*, meaning that $\mathcal{E}(u, v) = 0$ if v is constant on the support of u.

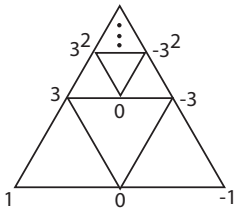

Figure 1.4.1

1.5 ELECTRIC NETWORK INTERPRETATION

In Section 1.3 we began by considering a notion of energy on a general finite graph G. More generally, suppose we have a positive function c_{xy} defined on the edges $x \sim y$ of the graph (we call this a *network*). Then we may consider

$$E(u) = \sum_{x \sim y} c_{xy}(u(x) - u(y))^2 \tag{1.5.1}$$

as an associated energy. We will interpret c_{xy} as conductances and the reciprocals $r_{xy} = 1/c_{xy}$ as resistances. We imagine an electric network where each vertex of G is a node and the edges of G are resistors connecting the nodes with the given resistances. The values of u are interpreted as voltages at the nodes. A current of amperage $(u(x) - u(y))/r_{xy} = c_{xy}(u(x) - u(y))$ will flow through each resistor, producing an energy of $c_{xy}(u(x) - u(y))^2$ from each resistor, leading to the total energy (1.5.1). Note that we have to do something (such as attach appropriate strength batteries) to keep the nodes at the specified voltages $u(x)$.

We could also drop the reference to the graph structure of G and require $c_{xy} \geq 0$ to be given as a symmetric function on all distinct pairs x, y in V. We then define $x \sim y$ if and only if $c_{xy} > 0$. If $c_{xy} = 0$ then the resistance is infinite, and it won't change the network to connect x and y by an infinite resistor. When we define the restriction of a network in what follows, we are essentially using this approach to define the edge relation.

We might also consider what happens if we impose voltages $u(x)$ at only some of the nodes (V') and allow the voltages at the other nodes (V'') to settle into values that, according to electric network theory, will minimize the energy. For example, suppose the network has three nodes x, y, z and two edges $x \sim y$ and $y \sim z$. If we set voltages $u(x)$ and $u(z)$ at the extreme nodes, then the value $u(y)$ at the middle node that minimizes

$$c_{xy}(u(x) - u(y))^2 + c_{yz}(u(y) - u(z))^2 \tag{1.5.2}$$

is easily seen to be

$$u(y) = \frac{c_{xy}u(x) + c_{yz}u(z)}{c_{xy} + c_{yz}}, \tag{1.5.3}$$

and this yields the value

$$\left(\frac{c_{xy}c_{yz}}{c_{xy} + c_{yz}}\right)(u(x) - u(z))^2 = \frac{1}{r_{xy} + r_{yz}}(u(x) - u(z))^2 \tag{1.5.4}$$

for (1.5.2). Note that this is the same value as the energy for a network with two nodes x and z connected by a resistor of resistance $r_{xy} + r_{yz}$. This is a familiar rule: Resistors in series add their resistances. The rule that resistors in parallel add their conductances is more or less built into the energy formula (1.5.1).

It seems reasonable that whatever the choice of nodes V', we could construct a network on V' that mimics the energy on the original network for any choice of values $u(x)$ for $x \in V'$. Any such network will be called a *restriction* of the original network to V', and the original network will be called an *extension* of the network

on V'. We will not be concerned here with abstract existence and uniqueness theorems for restrictions, since in all cases of interest we will compute restrictions explicitly.

From this point of view, the solution of the renormalization problem relating the energies $\mathcal{E}_0$ and $\mathcal{E}_1$ on the graphs Γ_0 and Γ_1 for SG is the same as starting with a network on Γ_1 with all resistance equal to r and showing that the restriction to V_0 is the graph Γ_0 with all resistances equal to 1. Here we will re-derive the answer in a step-by-step fashion using four basic principles of electric network theory, applied to pieces of the graphs. We have already mentioned the first two, resistors in series and resistors in parallel. The others are "pruning" and the "$\Delta - Y$ transformation."

LEMMA 1.5.1 *Suppose the deleted vertices V'' are connected only to each other and to a single vertex x_0 in V'. Then the restriction network is obtained by retaining all the edges connecting nodes in V' with the same resistances, and the minimum energy function has $u(y) = u(x_0)$ for every $y \in V''$.*

Proof: Given u on V', the choice $u(y) = u(x_0)$ for all $y \in V''$ adds zero to the sum

$$\sum_{\substack{x\sim y\\ x,y\in V'}} c_{xy}(u(x)-u(y))^2$$

and so clearly minimizes energy. □

LEMMA 1.5.2 *Consider a Y-shaped network with nodes x, y, z, w and edges just connecting x, y, z to w, and resistances r_{xw}, r_{yw}, r_{zw}. Then the restriction to $V' = \{x, y, z\}$ is a Δ-shaped network with resistances r_{xy}, r_{yz}, r_{zx} provided that*

$$\begin{cases} r_{xw} = \frac{r_{xy}r_{zx}}{R}, \qquad r_{yw} = \frac{r_{xy}r_{yz}}{R}, \qquad r_{zw} = \frac{r_{yz}r_{zx}}{R} \\ \text{for} \quad R = r_{xy}+r_{yz}+r_{zx}. \end{cases} \tag{1.5.5}$$

Moreover, the energy-minimizing value is

$$u(w) = \frac{c_{xw}u(x)+c_{yw}u(y)+c_{zw}u(z)}{c_{xw}+c_{yw}+c_{zw}}. \tag{1.5.6}$$

In particular, if $r_{xy} = r_{yz} = r_{zx} = a$ then $r_{xw} = r_{yw} = r_{zw} = a/3$, and $u(w) = \frac{1}{3}(u(x)+u(y)+u(z))$.

Proof: The Y-network energy is

$$c_{xw}(u(x)-u(w))^2 + c_{yw}(u(y)-u(w))^2 + c_{zw}(u(z)-u(w))^2. \tag{1.5.7}$$

It is clear that to minimize this we must choose $u(w)$ by (1.5.6). When we substitute (1.5.6) into (1.5.7) and simplify we obtain

$$c_{xy}(u(x)-u(y))^2 + c_{yz}(u(y)-u(z))^2 + c_{zx}(u(z)-u(x))^2 \tag{1.5.8}$$

for certain coefficients. After some messy algebraic manipulations we obtain (1.5.5). The details are left to the exercises. The special case of equal resistances is easy. □

Now we analyze the restriction of the Γ_2-network for SG with equal resistances. To simplify the computation we set the resistances equal to 1, since the result is

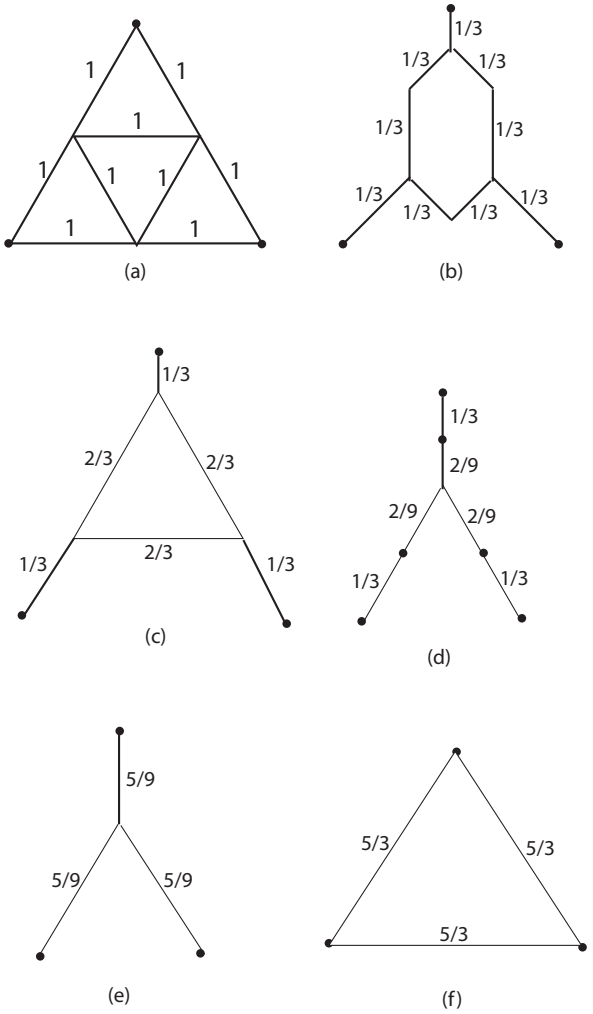

Figure 1.5.1

clearly linear in the resistance. We obtain the network in Figure 1.5.1(a), with the V_0 vertices marked by larger dots. We observe three Δ-shaped subnetworks, so we apply the $\Delta - Y$ transform to each, to obtain Figure 1.5.1(b). We see three sets of resistors in series, so we combine them to obtain Figure 1.5.1(c). Another Δ-shaped subnetwork appears, so we use $\Delta - Y$ to obtain Figure 1.5.1(d). Again there are three sets of resistors in series, so we combine to obtain the Y-shaped network in Figure 1.5.1(e). Finally, we do the inverse of the $\Delta - Y$ transform to obtain the network on V_0 in Figure 1.5.1(f). Keeping track of the resistances along the way, we see that we have multiplied by $\frac{5}{3}$, so if we started with $r = \frac{3}{5}$ we would end up with resistance 1. This confirms our calculation of the energy renormalization factor.

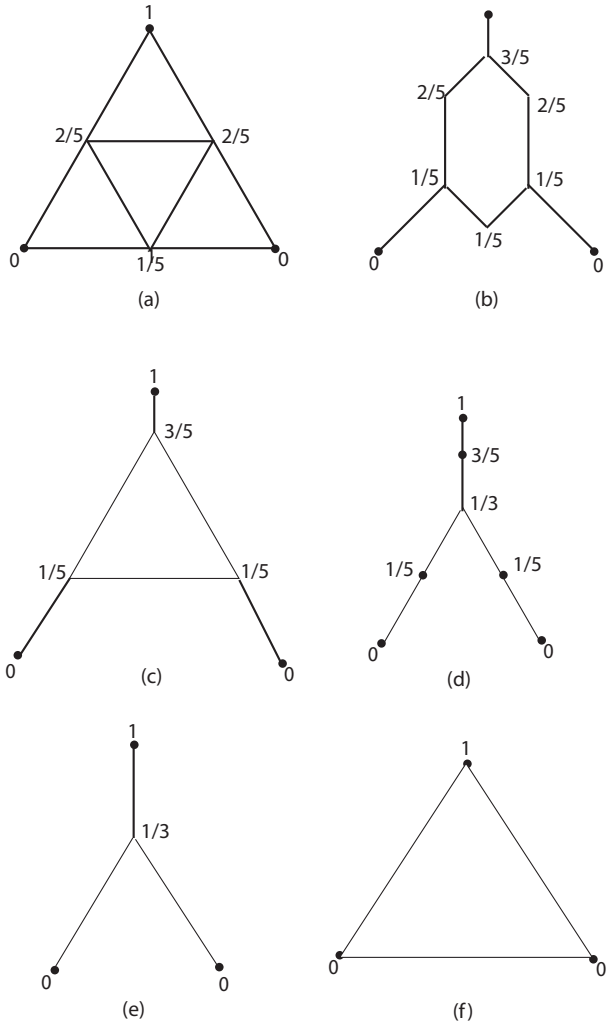

Figure 1.5.2

To obtain the harmonic extension algorithm we start with values of u at the vertices in Figure 1.5.1(f) and run the process backwards, using (1.5.3) and (1.5.6) to fill in the values when we add vertices. We show the results in Figure 1.5.2. For simplicity we just do the function h_0, but by symmetry the analogous computation holds for h_1 and h_2, and by linearity for all harmonic functions. Thus we rediscover the "$\frac{1}{5} - \frac{2}{5}$ rule".

Which method is easier? The first method involves solving a system of linear equations. The second method gives a step-by-step procedure that leads to the solution. Of course, Gaussian elimination also gives a step-by-step procedure that leads to the solution, but it is not the same procedure. In this case the system

of equations was easy to solve, so the first method was simpler. In other examples it seems that network manipulation is simpler. But not all networks allow simiplification using the few rules we have at hand. So, from a practical point of view, network manipulation is a tool to be used sparingly, though sometimes to great effect. From a theoretical point of view, it motivates and explicates the notion of effective resistance metric that we discuss in the next section.

EXERCISES

1.5.1.* Show that the restriction of a network to a subset $V' \subseteq V$ defines a unique network on V'.

1.5.2. Suppose $V''' \subset V' \subseteq V$. Given a network on V, show that its restriction to V''' is equal to the restriction to V''' of the restriction to V'.

1.5.3. Supply the details of the proof of Lemma 1.5.2.

1.5.4. Show that a network is connected if and only if the only functions of zero energy are the constants. Use this to show that the restriction of a connected network is connected.

1.5.5. Let $x_1 < x_2 < \cdots < x_n$ be points on the line, and define a network with $x_j \sim x_{j+1}$ and $r_{x_j x_{j+1}} = x_{j+1} - x_j$. Determine the restriction of this network to any subset of $\{x_j\}$.

1.5.6. Invert (1.5.5) to express r_{xy}, r_{yz}, r_{zx} in terms of r_{xw}, r_{yw}, r_{zw}.

1.6 EFFECTIVE RESISTANCE METRIC

Given any network, we can define the effective resistance $R(x, y)$ between any two points as the resistance between them when we restrict the network to just those two points. This is exactly the resistance we would measure if we attached a battery to the two points and measured the current flow. It should not be confused with the resistance of an edge connecting x and y. Such an edge need not exist! If we look at the definition of restriction of networks, we find the following formula:

$$R(x, y)^{-1} = \min\{\mathcal{E}(u) : u(x) = 0 \quad \text{and} \quad u(y) = 1\}. \tag{1.6.1}$$

Another formulation is that $R(x, y)$ is the minimum value of R such that

$$|u(x) - u(y)|^2 \le R\mathcal{E}(u) \quad \text{for all } u \in \operatorname{dom}\mathcal{E}. \tag{1.6.2}$$

We note that the function achieving the minimum in (1.6.1) is the function that is harmonic in the complement of the points x and y. For example, if the network is Γ_m for $K = I$, then (assuming $x < y$) the function u is 0 on $[0, x]$, 1 on $[y, 1]$, and linear on $[x, y]$. Clearly

$$\mathcal{E}_m(u) = \mathcal{E}(u) = \int_x^y \frac{1}{(y-x)^2}\, dt = \frac{1}{y-x},$$

so $R(x, y) = y - x$, the usual distance on I.

This leads us to hope that effective resistance will provide us with a natural (or intrinsic) metric on SG. What do we mean by this? The simplest interpretation is to first define $R(x, y)$ for points in V_m using the network Γ_m. Then we can extend it to V_*, since it clearly is independent of m (once m is large enough that $x, y \in V_m$). It is not difficult to see that $R(x, y)$ is uniformly continuous in x and y, so we may extend it to $SG \times SG$, and in fact (1.6.1) still holds. But is it a metric?

The claim is that effective resistance is a metric for any network. The only nontrivial condition to check is the triangle inequality. So given three points x, y, z, consider the restriction of the network to $\{x, y, z\}$. By Exercise 1.5.2, the effective resistances will be the same if we compute them with respect to this three-point network. So we only have to check the triangle inequality for a Δ-network. (There is also the trivial case when only two of the resistances are finite, where the triangle inequality is an equality.) This is an easy exercise, but it becomes quite obvious by doing a $\Delta - Y$ transformation. On the Y-network, $R(x, y) = r_{xw} + r_{yw}$, and so on, so

$$R(x, y) + R(y, z) = r_{xw} + r_{yw} + r_{yw} + r_{zw} > r_{xw} + r_{zw} = R(x, z).$$

Returning to SG, we know that effective resistance is a metric on V_m, hence on V_*, and by continuity on SG. We will soon see that it defines the same topology as the Euclidean metric, but it is not metrically equivalent. We note that (1.6.2), which may be written

$$|u(x) - u(y)| \le \mathcal{E}(u)^{1/2} R(x, y)^{1/2} \quad \text{for all } u \in \operatorname{dom}\mathcal{E}, \tag{1.6.3}$$

says that functions on $\operatorname{dom}\mathcal{E}$ are Hölder continuous of order $\frac{1}{2}$ in the effective resistance metric.

It is extremely difficult to compute $R(x, y)$, but it is rather easy to obtain approximate values. First we note that if we can construct any function u satisfying $u(x) = 0$ and $u(y) = 1$, then this immediately gives us the lower bound

$$R(x, y) \ge \mathcal{E}(u)^{-1}. \tag{1.6.4}$$

To find an upper bound we need to show that $u(x) = 0$ and $u(y) = 1$ implies $\mathcal{E}(u) \ge a$, as this implies

$$R(x, y) \le a^{-1}. \tag{1.6.5}$$

In particular, suppose $x, y \in V_m$ are neighboring vertices. Now we choose $u = \psi_y^{(m)}$, the piecewise harmonic spline in $S(\mathcal{H}_0, V_m)$ with $\psi_y^{(m)}(z) = \delta_{yz}$ for $y, z \in V_m$. Then we have $u(x) = 0$ and $u(y) = 1$, and $\mathcal{E}(u) = 4r^{-m}$ (or $2r^{-m}$ if y is a boundary point). So (1.6.4) says

$$R(x, y) \ge \frac{1}{4} r^m. \tag{1.6.6}$$

On the other hand, $u(x) = 0$ and $u(y) = 1$ implies $\mathcal{E}(u) \ge \mathcal{E}_m(u) \ge r^{-m}$, so (1.6.5) says

$$R(x, y) \le r^m. \tag{1.6.7}$$

Together, the two estimates show that $R(x, y) \approx r^m$. The same reasoning extends to other points.

LEMMA 1.6.1 *There exist positive constants c_1 and c_2 such that*
(a) if x and y belong to the same or adjacent m-cells, then

$$R(x, y) \le c_1 r^m; \tag{1.6.8}$$

(b) if x and y do not belong to the same or adjacent m-cells, then

$$R(x, y) \ge c_2 r^m. \tag{1.6.9}$$

Proof: In case (a) we construct chains of points joining x and y as in the beginning of Section 1.4. Using the triangle inequality and estimate (1.6.7) for pairs of consecutive points in the chain, by summing a geometric series we obtain (1.6.8). In case (b), let z_0, z_1, z_2 denote the boundary points of an m-cell containing y. Then $u = \psi^{(m)}_{z_0} + \psi^{(m)}_{z_1} + \psi^{(m)}_{z_2}$ is identically 1 on the m-cell containing y, but $u(x) = 0$. Also $\mathcal{E}(u) = 6r^{-m}$ (or $4r^{-m}$ if the cell intersects V_0), so (1.6.4) implies (1.6.9). □

It is easy to see that this means

$$R(x, y) \sim |x - y|^\beta \quad \text{for } \beta = \log \frac{5}{3} / \log 2. \tag{1.6.10}$$

This shows that the resistance metric is topologically equivalent, but not metrically equivalent, to the Euclidean metric. Since $\beta < 1$, it follows that distances in the resistance metric are much larger than in the Euclidean metric. In particular, there are no rectifiable curves in this metric. It seems very unlikely that SG in this metric can be embedded in a Euclidean space of any dimension.

There is no exact scaling identity relating $R(x, y)$ and $R(F_i x, F_i y)$. We can say that, roughly speaking, each F_i acts like a contraction of ratio r.

EXERCISES

1.6.1. Show the equivalence of (1.6.1) and (1.6.2).
1.6.2. Prove that $R(x, y)$ is uniformly continuous on $V_* \times V_*$ in SG, and (1.6.1) holds on all of $SG \times SG$.
1.6.3. Compute the effective resistance on a Δ-network directly (without using the $\Delta - Y$ transform), and show that it is a metric.
1.6.4. Compute $R(x, y)$ exactly on $V_1 \times V_1$.
1.6.5. Give the details of the proof of Lemma 1.6.1(a).
1.6.6. Prove (1.6.10) from Lemma 1.6.1.
1.6.7. Show that $\mu\{x : R(x, y) \le r\} \sim r^d$ for $d = \frac{\log 3}{\log(5/3)}$, where μ is the standard measure. This means that SG as a metric-measure space, with metric R and measure μ, has dimension d.

1.7 NOTES AND REFERENCES

Most of the material in Sections 1.1, 1.3, and 1.4 is from [Kigami 1989], where it was developed for SG and its higher dimensional analogs. The fact that it also has something to say about I is a pleasant afterthought. Although it doesn't say

anything new, it gives an amazingly simple characterization of the Sobolev space H^1; no Lebesgue integration theory or Schwartz distribution theory is needed, just a plain calculus-type limit. Of course, to show the equivalence of H^1 and $\operatorname{dom}\mathcal{E}$ in Exercise 1.4.6 requires all this machinery.

The definition of self-similar measure in Section 1.2 is from [Hutchinson 1981]. We take a very naive approach to measures in general because we only have to integrate continuous or piecewise continuous functions, so we can use a Riemann-type integral. The positivity (1.2.1) and continuity (1.2.3) conditions are not part of the usual definition of measure, and we have made the ad hoc definition of "regular" to describe them. Occasionally we have to consider more general measures in the sequel.

The topological rigidity of SG (Exercise 1.1.6) was first noted in [Bandt and Retta 1992]. The fact that $\widetilde{SG}$ is not topologically rigid (Exercise 1.1.7) can also be understood from the Apollonian packing model. See [Mumford et al. 2002] for beautiful pictures of this.

Exercise 1.3.5 is a warmup for the singularity of energy measures [Kusuoka 1989]. Exercise 1.3.9 is from [Stanley et al. 2003]. The singular harmonic function in Exercise 1.4.7 was first noted in [Dalrymple et al. 1999].

The electric network ideas in Sections 1.5 and 1.6 come from [Kigami 1994a]. See [Doyle and Snell 1984] for the general theory of networks. See [Fukushima et al. 1994] for the general theory of Dirichlet forms.

It is possible to identify $\operatorname{dom}\mathcal{E}$ with a certain Lipschitz-type function space determined by the embedding of SG in the plane, as shown in [Jonsson 1996]. This is one result that contradicts my assertion that the standard embedding of SG in the plane is irrelevant for our analytic theory. Nevertheless, it seems to be a kind of isolated result, since other natural function spaces on SG are unrelated to the embedding [Strichartz 2003b].