Figure 3.4. Voronoi decomposition of a surface with a non-Euclidean metric.

and Letscher [248], showing the existence of a Voronoi tessellation of a general Riemannian manifold sampled with "sufficient density." To give a quantitative definition of what is meant by "sufficient density," the authors resort to the notion of convexity radius. Recall that in Chapter 2, we defined a convex set in X as a set in which the minimal geodesic between each pair of points lies inside the set. The *convexity radius* of the surface X at a point x is the largest r for which the closed ball $\overline{B}_r(x)$ is convex in X. The convexity radius of the entire X is simply the infimum of the convexity radii over all points in X. Informally, we may say that on the scale below the convexity radius, the surface behaves very much like a Euclidean space – any interesting topological and geometric properties, like the one in Figure 3.5, appear on larger scales. Leibon and Letscher prove that if X' is an r-separated r-covering with r smaller than $\frac{1}{5}$ the convexity radius of X, then the Voronoi regions are topological disks and the Voronoi decomposition generated by X' is a valid tessellation of the surface [248, 301].

3.4 Centroidal Voronoi sampling and the Lloyd-Max algorithm

The notion of Voronoi tessellation allows us to express the sampling of a surface as a mapping $y^* : X \to X'$, copying the interior of each V_i to its *closest point* x_i in X' (the boundaries shared by more than one Voronoi cell can be copied to x_i belonging to any of the intersecting cells). Thinking of sampling in these terms, it is natural to quantify the error introduced by

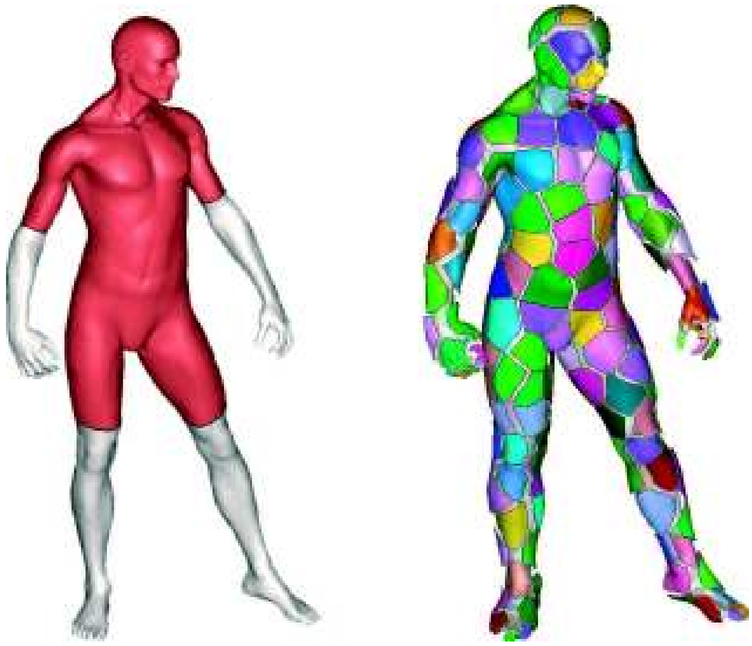

Figure 3.5. Left: Voronoi decomposition of a surface with insufficient sampling density. The shaded Voronoi region is not homeomorphic to a disk. Right: increasing the sampling density produces a valid tessellation. Observe that now the body is decomposed into topological disks.

replacing x with its representation $y^*(x)$. For that purpose, let us assume x is chosen at random with uniform distribution on X, where by *uniform distribution* we mean that the probability that x falls into a subset A on X is proportional to the area of A. Formally, this can be written as

$$\mathbb{P}(x \in A) = \frac{\mu(A)}{\mu(X)} = \frac{1}{\mu(X)} \int_A da,$$

where $\mu(X)$ is the area of X, da is the differential area element, and $\mathbb{P}$ stands for probability. The representation error associated with a sampling X' can be expressed as the *variance* of the random variable $d_X(x, y^*(x))$,

$$\begin{aligned} \varepsilon(X') &= \mathrm{Var}(d_X(x, y^*(x))) \\ &= \frac{1}{\mu(X)} \int_X d_X^2(x, y^*(x)) = \frac{1}{\mu(X)} \sum_{i=1}^{N} \int_{V_i(X')} d_X^2(x, x_i) da. \end{aligned} \quad (3.2)$$

In the Euclidean case, the latter expression becomes

$$\varepsilon(X') = \frac{1}{\mu(X)} \sum_{i=1}^{N} \int_{V_i(X')} \|x - x_i\|_2^2 dx,$$

which can be interpreted as the *mean squared error* of the representation.

The representation error $\varepsilon(X')$ gives a quantitative measure of the sampling quality and leads to the natural question of finding the best sampling. Formally, we are looking for a set X' of N points on X, bringing the error $\varepsilon(X')$ to minimum among all sets of N points on X. A related problem is finding the smallest sampling X' having $\varepsilon(X')$ below some predefined value. It appears that a similar question arises in a variety of fields. For example, in image processing, a continuous domain of vectors encoding the intensities of different colors in an image often need to be represented by a finite set of symbols. Computation of such a representation is known as *vector quantization* [185, 172]. In machine learning, pattern recognition, and data mining, we encounter objects represented as vectors of features, and it is often required to aggregate them into groups sharing some common trait. The process of partitioning a space into such groups is termed *clustering* or *unsupervised learning* [141].

It can be shown (see Problem 3.5) that in order for a sampling X' to minimize $\varepsilon(X')$, each point x_i has to satisfy

$$x_i = \arg\min_{x \in V_i} \int_{V_i} d_X(x, x') da.$$

A point minimizing[4] the latter integral is called the *intrinsic centroid*[5] of V_i [140]. To understand the origin of this name, note that in the Euclidean case the condition becomes

$$x_i = \arg\min_{x \in V_i} \int_{V_i} \|x - x'\|_2 dx' \;=\; \frac{\displaystyle\int_{V_i} x\, dx}{\displaystyle\int_{V_i} dx},$$

which is simply the *centroid* (or *center of mass*) of the set V_i. A Voronoi tessellation generated by a set of points x_i, which are themselves the intrinsic centroids of the corresponding Voronoi cells V_i, is called a *centroidal Voronoi tessellation*. We refer to a sampling associated with a centroidal Voronoi tessellation as to a *centroidal Voronoi sampling*. Such a sampling is optimal in the sense of $\varepsilon(X')$ and is not unique.

A simple way to compute a centroidal Voronoi sampling starts by picking up an arbitrary sampling X' (produced, for example, using the farthest point sampling algorithm), and computing the associated Voronoi tessellation. Next, we compute the intrinsic centroids for each Voronoi cell and use them as a new sampling X'. Clearly, the Voronoi tessellation has changed and needs to be recomputed. Repeating this process several times gives a reasonable approximation of a true centroidal Voronoi tessellation. The entire procedure can be summarized as shown in Algorithm 3.2.

input : metric space (X, d_X), initial set of points X'.
output : optimal set of points X', minimizing $\epsilon(X')$.
repeat
    Construct the Voronoi tessellation associated with X'.
    Compute the intrinsic centroids of V_i and set X' to be these points.
until X' *stops changing significantly*

Algorithm 3.2. Lloyd-Max algorithm.

This procedure is known as *Lloyd-Max algorithm*[6] in signal and image processing [255, 263, 256] and *k-means* in statistics [260]. We are not going to explore the full details of this algorithm and refer the reader to [312] for additional information. We only mention that Lloyd-Max algorithm is a very simple *alternating minimization* procedure, which attempts to produce a sequence of samplings that decrease the value of $\varepsilon(X')$ [107]. We will encounter more sophisticated numerical recipes for minimization of functions in Chapter 5.

Compared with the greedy farthest point sampling, which never changes the locations of the points previously added to the sampling and optimizes only the location of the next point, centroidal Voronoi sampling allows us to change the locations of all the points. As a consequence, the produced sampling is more uniform and may require less samples than one produced by FPS. In fact, as the number of samples grows asymptotically, all Voronoi cells converge to a hexagonal shape, which is known to produce the densest possible tessellation ("honeycomb" tiling, called this way because it is often encountered in Nature, see Figure 3.6). This result was known for a while in the Euclidean case and has been proved only recently for two-dimensional Riemannian manifolds [191].

Figure 3.6. Nature surprises us by producing a variety of living examples of Voronoi tessellations, including the spot-shaped coloring of an African giraffe (left), the pattern on the shell of a *testudo hermanni* turtle (middle), and the hexagonal wax honeycomb cells built by honey bees (right).