Sampling from curves, surfaces, and other coordinatized regions

Dhiren Ghosh
Synectics for Management Decisions
1901 North Moore Street, Suite 900
Arlington, VA 22209 U. S. A.

Andrew Vogt
Department of Mathematics
Georgetown University
Washington, DC 20057-1233 U. S. A.

1 Introduction

A problem that arises in geographical and geological surveys, the physical and biological sciences, computer graphics, and quality control is how to select a random sample from a constrained set such as a curve in two- or three-dimensional space or a surface or body in three-dimensional space. We show here how to solve this problem provided the set can be coordinatized in a reasonably regular way. We discuss how to choose points on curves, surfaces, and higher-dimensional manifolds. We provide examples and note some related issues.

First of all, a caveat is in order. Here we identify a random sample as one chosen on the basis of equiprobability (and independence). However, selecting a finite sample from an infinite population so that each element of the population is equally likely to be chosen is not a meaningful task. The only way the elements can have equal probability is if each has probability zero. To avoid this difficulty, we divide the population into finitely many nonoverlapping regions of equal measure (say equal length, area, or volume) so that all of the points in a given region are indistinguishable for practical purposes. Then we pick one of the regions at random and any point in that region.

If the total measure of the population is infinite, we face another difficulty, though, namely, that the individual regions also have infinite measure. Usually points in a region of infinite measure are quite distinguishable from one another. In addition, the reason for requiring that the regions be of equal measure is that our notion of randomness is based on the measure: regions of equal measure are presumed to be equiprobable. But this criterion is inadequate when applied to

sets of infinite measure because such sets typically can be decomposed into two or more disjoint subsets also of infinite measure. If a set and two of its disjoint subsets are assumed to have the same probability p, then $p + p \leq p$ and $p = 0$. To choose a random sample from a population, not only should we partition the population into finitely many regions of practically indistinguishable points, each region of equal measure, but also we should require that the total measure be finite so that the regions may plausibly be treated as equiprobable. Furthermore, points within regions of small measure may be quite dispersed and thus quite distinguishable from one another. For example, the rational numbers are dense in the reals or in any finite interval of reals, and they form a set of measure zero. However, rational numbers, for example, integers, may be quite distinguishable from one another

The issues can be illustrated by the problem of selecting a number randomly from an interval of finite length, say the unit interval $[0, 1]$. The usual solution to this problem is to consult a random number table, i.e., choose k digits $x_1, x_2, ..., x_k$ randomly and independently from $0, 1, ..., 9$ to obtain the number $\frac{x_1}{10} + \frac{x_2}{10^2} + ... + \frac{x_k}{10^k}$. This selection is actually made from the finite set of fractions with denominator 10^k rather than from the set $[0, 1]$. Alternatively, the selection can be viewed as the choice of a subinterval from the interval $[0, 1]$, with the understanding that it is immaterial which point is actually chosen from within the selected subinterval, each of the 10^k possible subintervals having the same length $\frac{1}{10^k}$.

In the following we propose to consider only cases where the curve, surface, etc. has finite total measure, subsets of equal measure are equiprobable, and the decomposition is into finitely many regions of equal measure such that points within a region are indistinguishable for practical purposes. The last condition, we will assume, is realized when the parameters of points within a given region are close to each other under a regular [1] parametrization of the curve, surface, etc.

2 Sampling on curves

Consider a plane curve $\{(x(t), y(t)) : a \leq t \leq b\}$, where $t \to x(t)$ and $t \to y(t)$ are continuously differentiable functions (with one-sided derivatives at the two endpoints) and such that $t \to (x(t), y(t))$ is one-to-one for $a \leq t < b$. To select a random sample on such a curve, we decompose the t-interval $[a, b]$ into subintervals so that t-subintervals have weights proportional to the arc length they sweep out along the curve. The measure of arc length is $ds = \frac{ds}{dt} = \sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}$.

Lahiri's method (see [Cochran, 1977, p. 251]) can be used to perform the selection. Let $M = \max\{\frac{ds}{dt}(t) : a \leq t \leq b\}$. Choose pairs of points (t, u) randomly from the rectangle $[a, b] \times [0, M]$, i.e., choose t randomly from the in-

[1] A regular parametrization is one in which the Jacobian has maximal rank except on a set of measure zero in the parameter set.

terval [a,b], and u randomly and independently from the interval [0,M]. Then omit points (t,u) for which u exceeds $\frac{ds}{dt}(t)$. Then calculate $(x(t), y(t))$ for the t-values that remain, and these are the desired points. More accurately we choose a t-subinterval and a u-subinterval, and thus a small coordinate rectangle in $[a, b] \times [0, M]$, pick a point (t,u) arbitrarily in this rectangle (all points in the rectangle judged to be practically equivalent), and determine whether $\frac{ds}{dt}(t)$ is less than u. Even this final determination is only approximate since the computation of $\frac{ds}{dt}(t)$ will involve some rounding.

A piecewise continuously differentiable curve or a curve with isolated self-intersections assumed to have finite total length can be decomposed into a finite number of curves of the type in the previous paragraph. If these curves have lengths $L_1, L_2, ..., L_k$ with total length $L = \sum_{i=1}^{k} L_i$, we pick a number z randomly from the interval [0,L] and if $z \leq L_1$, we apply Lahiri's method to the first curve, or if $L_1 < z \leq L_1 + L_2$ we apply the method to the second curve, and so on [2].

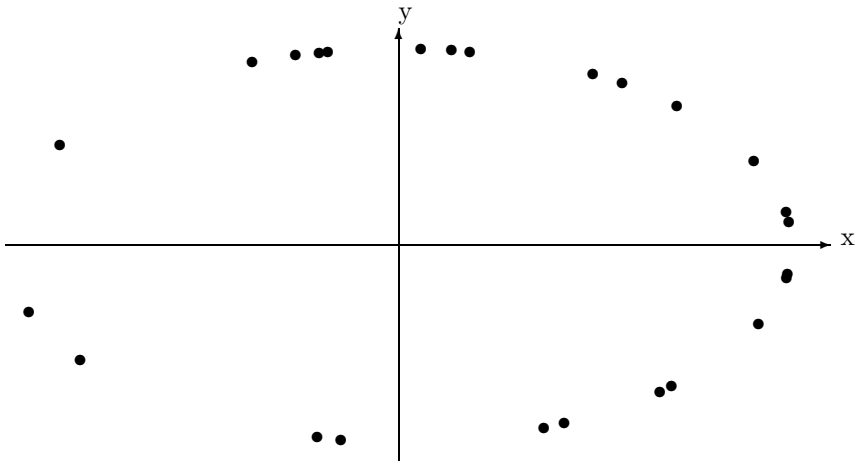
Consider for example random selection from the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ with $a \geq b > 0$. A standard parametrization for this curve is $t \to (x(t), y(t)) = (a\cos t, b\sin t)$ for $0 \leq t \leq 2\pi$. Then $\frac{ds}{dt}(t) = \sqrt{(-a\sin t)^2 + (b\cos t)^2}$, and the latter has maximum value a. So choose a pair of numbers (t,u) uniformly in the rectangle $[0, 2\pi] \times [0, a]$ and let the sample point be $(a\cos t, b\sin t)$ provided $u \leq \sqrt{a^2\sin^2 t + b^2\cos^2 t}$. (Choosing "uniformly" is a misnomer because, as we mentioned above, the method of selection is ultimately finitary ands depends on a particular discretization, but we shall continue to use the traditional language.)

To show even more detail, suppose our ellipse is $\frac{x^2}{4} + y^2 = 1$. Then we pick a random pair of numbers in $[0, 2\pi] \times [0, 2]$, say, (1.774, 1.291). When we compute $\frac{ds}{dt}(t) = \sqrt{4\sin^2 t + \cos^2 t} = \sqrt{3\sin^2 t + 1}$ at t = 1.774, we get: $\frac{ds}{dt} \approx 1.969 > u = 1.291$. Thus we choose the point $(x, y) = (2\cos 1.774, \sin 1.774) \approx (-.404, 0.979)$.

t	u	$\frac{ds}{dt}(t)$	(x(t),y(t))
1.774	1.291	≈ 1.969	$\approx (-.404, 0.979)$

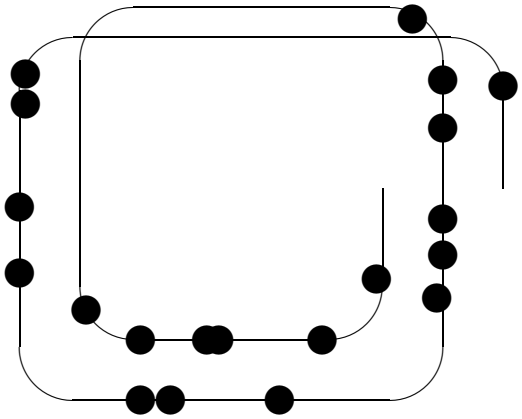
We repeat this a total of thirty-one times, generating a total of twenty-five points and rejecting seven points. As can be seen from the figure, the points obtained are not evenly distributed around the ellipse. However, this is the nature of simple random sampling.

[2]Lahiri's method can be used to select the curve: choose i equiprobably from 1, 2, ..., k and v randomly from the interval $[0, \max\{L_1, L_2, ..., L_k\}$, and keep i provided $v \leq L_i$.



The same method, mutatis mutandis, can be applied to curves in any dimension with $t \to (x_1, ..., x_n)$ replacing $t \to (x, y)$.

Consider, for example, the helix $(x, y, z) = (\cos t, \sin t, \frac{2}{\pi}t)$ for $0 \leq t \leq 4\pi$. We choose t randomly from the interval [0, 4π]. Since $\parallel (\frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt}) \parallel = \sqrt{\sin t^2 + \cos t^2 + \frac{4}{\pi}} \equiv \sqrt{1 + \frac{4}{\pi}} = M$, we choose the second variable u randomly in [0,M] and keep the point with parameter value t provide $u \leq M$. But this is always true. So we do not need u. In the Figure below we show twenty points so obtained (because of the publication requirement of embedded graphics, our helix has been approximated by a series of ovals in LaTeX.)



3 Sampling on surfaces

In the case of a surface, we consider a parametrization of the form $\{X(u,v) : (u,v) \in D\}$ where D is an open bounded subset of $\mathcal{R}^2$ and $X(u,v) = (x(u,v), y(u,v), z(u,v)$ is a continuously differentiable function. The surface area element is:

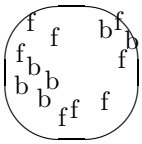
$$dA = \|\frac{\partial X}{\partial u} \times \frac{\partial X}{\partial v}\| \, dudv = \sqrt{(\frac{\partial(x,y)}{\partial(u,v)})^2 + (\frac{\partial(y,z)}{\partial(u,v)})^2 + (\frac{\partial(z,x)}{\partial(u,v)})^2} \, dudv.$$

The squared terms are two by two determinants obtained from the partial derivatives of x, y, and z with respect to u and v. Cf. [Marsden et al., 1993, Chapter 6]. If we wish to choose a point randomly from this surface, we choose a point (u,v,w) uniformly from a rectangular solid $(a,b) \times (c,d) \times (0,M)$ where $(a,b) \times (c,d)$ contains D and $M = \sup\{\|\frac{\partial X}{\partial u} \times \frac{\partial X}{\partial v}\| : (u,v) \in D\}$, the latter assumed to be finite. Then start over if (u,v) is not in D or if $w > \|\frac{\partial X}{\partial u} \times \frac{\partial X}{\partial v}\|$. Otherwise the point $X(u,v)$ is selected for the sample.

The simplest example of this kind of sampling is on the unit sphere $\{(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi) : 0 \le \phi \le \pi, 0 \le \theta \le 2\pi\}$. Then $\|\frac{\partial X}{\partial \phi} \times \frac{\partial X}{\partial \theta}\| = \|(\sin\phi)X\| = \sin\phi$, and hence M can be taken to equal 1. Now we pick points (ϕ,θ,w) uniformly in $(0,\pi) \times (0,2\pi) \times (0,1)$, discard the points if $w > \sin\phi$, and retain the points $(\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi)$ otherwise.

ϕ	θ	w	$\sin\phi$	x	y	z
0.132	3.007	0.215	0.132	-	-	-
2.416	5.648	0.547	0.663	0.534	-0.394	-0.748
2.707	2.905	0.594	0.421	-	-	-
1.440	0.501	0.201	0.992	0.870	0.476	0.130
1.007	0.971	0.053	0.845	0.477	0.698	0.534
0.567	3.472	0.274	0.537	-0.508	-0.174	0.844
2.564	2.145	0.225	0. 546	-0.296	0.459	-0.838
2.236	4.457	0.302	0.787	-0.200	-0.761	-0.617
1.616	2.365	0.640	0. 999	-0.713	0.700	-0.045
2.304	0.162	0.740	0.743	0.733	0.120	-0.669
1.577	0.935	0.704	1.000	0.594	0.804	-0.006
1.070	3.159	0.163	0.877	-0.815	-0.323	0.480
3.077	2.552	0.909	0.065	-	-	-
0.771	3.255	0.346	0.697	-0.692	-0.079	0.717
0.771	3.997	0.073	0.697	-0.457	-0.526	0.717
2.427	4.775	0.399	0.655	0.041	-0.654	-0.755
2.839	4.707	0.729	0.298	-	-	-
2.986	1.145	0.987	0.155	-	-	-
1.904	2.862	0.307	0.945	-0.908	0.261	-0.327
2.887	3.501	0.733	0.252	-	-	-

Below we see a rough depiction of where the points are on the sphere, with points in the front hemisphere labeled "f" and points in the back labeled "b".



Another way to sample points randomly from the unit sphere is to choose points (x, y, z) randomly from a cube, say. $[-1, 1] \times [-1, 1] \times [-1, 1]$, centered on the origin $(0, 0, 0)$, retain those points for which $x^2 + y^2 + z^2 \leq 1$, and project to the unit sphere by dividing the coordinates by $\sqrt{x^2 + y^2 + z^2}$. This method only works in special cases (for example, circles and regular polygons

in two dimensions; and spheres and the five Platonic solids and some semi-regular polyhedra in three dimensions; with projection from the center). For an ellipsoid such as $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$, where $a \le b \le c$ and not all three are equal, this method will not work, but our method does, with elliptical coordinates $\{(a \sin\phi \cos\theta, b \sin\phi \sin\theta, c \cos\phi) : 0 \le \phi \le \pi, 0 \le \theta \le 2\pi\}$ and $M = bc$.

For our next example we consider a torus. A standard parametrization is $\{((R+r\cos\phi)\cos\theta, (R+r\cos\phi)\sin\theta, r\sin\phi) : 0 \le \phi, \theta \le 2\pi\}$, where $R > r > 0$. Then $\|\frac{\partial X}{\partial\theta} \times \frac{\partial X}{\partial\phi}\| = \|((R + r\cos\phi)r\cos\theta\cos\phi, (R + r\cos\phi)r\sin\theta\cos\phi, (R + r\cos\phi)r\sin\phi)\| = (R + r\cos\phi)r$, and this quantity lies between 0 and $(R + r)r$. We select points (θ, ϕ, w) with θ, ϕ, and w uniform and independent on $[0, 2\pi]$, $[0, 2\pi]$, and $[0, (R + r)r]$. Our sample points are those points $X(\theta, \phi)$ for which $w \le (R+r\cos\phi)r$. Using Quattro Pro and MSPaint, we have depicted a sample of fifty points on the torus. Unfortunately the graphics cannot be embedded in our LaTeX file, and thus are omitted here.

4 Sampling on coordinatized regions

In higher dimensional cases we suppose that the region is a k-dimensional manifold in R^n parametrized by the mapping $u = (u_1, ..., u_k) \longmapsto X(u) = (x_1(u), ..., x_n(u))$ for u in an open bounded subset U of R^k and $X(u)$ a continuously differentiable function with bounded first derivatives. The volume element in the manifold is:

$$dV_k = \sqrt{\sum_{\{(i_1,...,i_k):1\le i_1<i_2...<i_k\le n\}} (\frac{\partial(x_{i_1}, ..., x_{i_k})}{\partial(u_1, ..., u_k)})^2} du_1...du_k = \rho(u)du_1...du_k,$$

where the squared terms are k by k subdeterminants of the matrix of partial derivatives $\{\frac{\partial x_i}{\partial u_j}\}$ [Schreiber, 1977, Appendix]. To select random points in this manifold, first pick u uniformly in U (or in a k-dimensional rectangular region that encloses U, retaining u only if it is in U) and then pick w uniformly and independently in $[0, M]$ where $M = \sup\{\rho(u) : u \in U\}$. Finally our random point is $X(u)$ provided $w \le \rho(u)$.

The k-dimensional volume of the above manifold is finite (since $\rho(u)$ is bounded) and is obtained by integrating dV_k over the parameter set U.

Sampling can also be done on a manifold that is a finite union of such k-dimensional manifolds (k fixed) with no intersections except in sets of lower dimension. If these manifolds have k-volumes $V_1, ..., V_m$, we pick an index i with probability $V_i / \sum_{j=1}^m V_j$ and then pick a point in the i-th manifold according to the scheme above.

5 Final remarks

The above methodology tells us how to pick a point randomly from a surface using a coordinatization of the surface and the natural length, area, or volume function, which plays the role of a probability density. Obviously this density could be replaced by another if we wished to choose points according to a different principle (e.g., picking more points where the curvature is greater).

Moreover, as our examples illustrate, randomly chosen points may be distributed far from evenly - with striking patterns and clusters visible in most samples. A variety of methods are available for improving the situation such as stratification and systematic sampling. Appropriate strata or systematics are easy to devise for curves. Generalizations for higher dimensions, though not always straightforward, have been obtained by the authors.

6 References

Cochran, W. (1977). *Sampling techniques.* Wiley, New York. Third edition.

Ghosh, D. and Vogt, A. (1997). An algorithm for sampling on surfaces. Bulletin of the International Statistical Institute, 51st session, Contributed Papers, Tome LVII, Book 1, pp. 629-630.

Marsden, J. E., Tromba, A. J., and Weinstein, A. (1993). *Basic Multivariable Calculus.* Springer-Verlag, New York.

Schreiber, M. (1977). *Differential Forms. A Heuristic Introduction.* Springer-Verlag, New York.