UNIFORM SAMPLING ON THE STANDARD SIMPLEX

ALLAN R. WILLMS

Abstract. Three methods for obtaining uniform sampling on the standard simplex are summarized and a new derivation of one method is provided, which only uses basic notions from calculus, difference equations, and differential equations.

Preprint of: Allan R. Willms, 2021, Missouri Journal of Mathematical Sciences, 33 (1), pp. 119-124.

1. Introduction

Obtaining a uniform sample of points from a unit cube (or a hypercube in $\mathbb{R}^n$) is an easy thing to do. One simply samples each coordinate from the interval $[0, 1]$. However, if the space to be sampled is not the hypercube but rather the standard simplex, Δ^n, defined by

$$\Delta^n = \left\{ x \in \mathbb{R}^{n+1} \;\middle|\; \sum_{i=1}^{n+1} x_i = 1, \;\; x_i \geq 0 \right\}, \qquad n \geq 0,$$

the way to sample this space uniformly is not so obvious. A naive approach would be to sample x_1 uniformly on $[0, 1]$, then sample x_2 uniformly on $[0, 1 - x_1]$, and so on, until x_n is obtained, and then set $x_{n+1} = 1 - \sum_{i=1}^{n} x_i$. This method will produce a point on the simplex, but, as illustrated in Fig. 1a for $n = 2$, repeated sampling by this means does not result in a uniform sampling of the simplex. On reflection, this method must obviously fail because x_1 has equal chance of being less than 0.5 or greater than 0.5, but the volume of the simplex in the region $x_1 < 0.5$ is clearly much larger than the corresponding volume with $x_1 > 0.5$. Another possible method would be to sample uniformly on the hypercube and scale each point back to the simplex, that is, for x in the unit hypercube, take $x/\sum x_i$ as a point on the simplex. However, this method also fails to give a uniform sampling of the simplex as illustrated in Fig. 1b. Again, upon reflection, it is easy to see that the corners of the simplex will be under-represented by such a sampling technique, since the volume of the hypercube that this scaling maps to those areas is small compared to that which maps to an equal size area of the simplex near its center.

Proper methods for uniform sampling on the standard simplex have been established for some time, although they are not necessarily well known outside the statistics community. Uniform sampling on Δ^n is closely related to the generation of ordered uniform samples used for many Monte Carlo simulations and is well-studied in that discipline.

The purpose of this short paper is twofold: to help provide awareness of the below three methods for uniform sampling on the simplex, and to provide a novel derivation of one of these methods, which uses only calculus rather than statistical theory. Further, a generalization for sampling on a simplex defined by $\sum_{i=1}^{n+1} c_i x_i = z > 0$, for some set of positive constants c_i is pointed out.

2. Three Methods

The text by Devroye [2] gives results useful for this problem in terms of uniform order statistics. If U_i, $1 \leq i \leq n$ are independent and identically distributed (iid) uniform random variables on $[0, 1]$ then the order statistics for this sample are these values sorted and labeled as $U_{(1)}, \ldots, U_{(n)}$, where

$$U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}.$$

The spacings are defined as $S_i = U_{(i)} - U_{(i-1)}$, $1 \leq i \leq n + 1$, where $U_{(0)} := 0$, and $U_{(n+1)} := 1$. Observe that the spacings add to one: $\sum_{i=1}^{n+1} S_i = 1$.

MSC2020: 62D05, 62-01
Key words and phrases: Standard simplex, uniform sampling

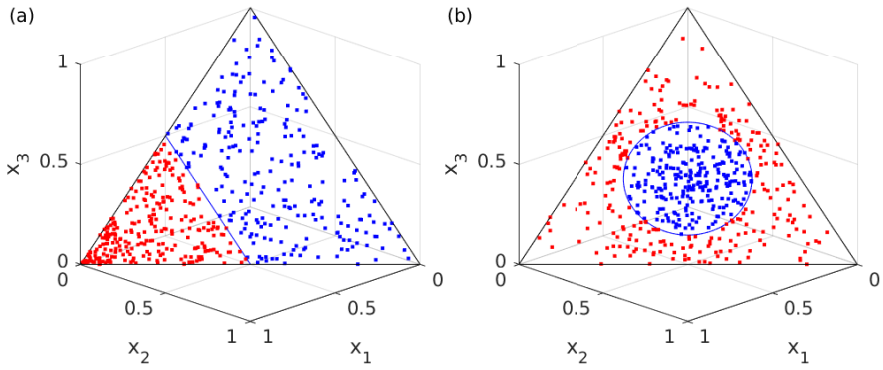

FIGURE 1. (a) Naive sampling, as described in the text, on the standard simplex in $\mathbb{R}^3$ does not produce a uniform sample of points. About half of the 500 generated points (blue) lie in the area $x_1 < 0.5$, but that area is three quarters of the total area of the simplex. (b) Uniform sampling on the cube and then scaling to the simplex Δ^2 also does not yield a uniform sample. The drawn circle, centered at $[1/3, 1/3, 1/3]$ contains half of the 500 points but corresponds to less than 27% of the total area of the simplex.

Devroye [2, Theorem 2.1 pg. 207] shows that the first n spacings are uniformly distributed over the region

$$A_n = \left\{ (x_1, \ldots, x_n) \;\middle|\; x_i \geq 0, \sum_{i=1}^{n} x_i \leq 1 \right\}.$$

This immediately leads to a method of uniformly sampling on Δ^n: obtain n uniform iid samples U_i on $[0, 1]$, sort them in order, $U_{(i)}$, and then the spacings S_i define a point that is uniformly distributed on Δ^n. For large n, the sorting step typically requires $\Theta(n \log n)$ operations, but Devroye [2, pg. 215] also gives a bucket sorting algorithm and a theorem exploiting the properties of the uniform distribution that shows that the expected time for the bucket sorting algorithm to complete is $O(n)$.

If one wishes to obtain a uniform sampling on Δ^n without sorting, Devroye [2, Theorem 2.2 pg. 208] shows that the $n + 1$ spacings S_i, $1 \leq i \leq n + 1$ are distributed as $E_1/G, \ldots, E_n/G$ where $E_1, \ldots, E_{n+1}$ are a sequence of iid exponential random variables, and $G = \sum_{i=1}^{n+1} E_i$. However, like the first method, this method requires two passes, in this case, the first to sample the random numbers E_i and the second to compute G and the spacings.

The third method Devroye [2, Theorems 2.3 and 2.4, pgs. 211–212] provides for producing order statistics for $U_1, \ldots, U_n$ is based on exponential spacings and two theorems he attributes to Sukhatme [9] and Malmquist [6]. The method is: set $U_{(n+1)} = 1$, then for i from n to 1, generate a uniform random variable f in [0,1] and set $U_{(i)} = f^{1/i} U_{(i+1)}$. From these order statistics, one can compute the spacings to get a uniform distribution on Δ^n, or equivalently one can modify the equation slightly to compute the spacings directly. The resulting formula is equivalent to the one presented in the next section. A minor advantage of this third method is that it is a one-pass method and, as pointed out by Bentley and Saxe [1], can be implemented as a subroutine to give one coordinate at a time without the need to generate all n random variables first. More importantly this method, like the second, does not require a sorting step.

Devroye attributes the last two methods to Lurle and Hartley [4], Schucany [8], and Lurle and Mason [5]. Devroye also cites the experimental comparison studies by Rabinowitz and Berenson [7], Gerontides and Smith [3], and Bentley and Saxe [1].

3. Calculus-Based Derivation of the Volume Splitting Method

The general idea for the method is to split the volume of the simplex in two by setting one coordinate to a constant, $x_i = y$, and then determining the relationship between y and the fraction ϕ, of the volume of the simplex that lies in the region $x_i \geq y$. A uniform sampling of the fraction ϕ can then be used to determine the corresponding value y for the coordinate x_i. The second major idea that makes the method

work particularly easily for the standard simplex is that after choosing one coordinate, the problem can be reduced to sampling on a simplex in one smaller dimension.

In order to derive the method, first generalize the definition of the simplex to allow the sum of the variables to be any fixed positive value z. Let

$$\Delta^n(z) = \left\{ x \in \mathbb{R}^{n+1} \;\middle|\; \sum_{i=1}^{n+1} x_i = z \;,\; x_i \geq 0 \right\}, \qquad n \geq 0. \tag{3.1}$$

The goal is to obtain uniform sampling on $\Delta^n(1)$.

A result from multivariable calculus is that the area of a surface defined implicitly by the equation $f(x_1, x_2, x_3) = 0$ for (x_1, x_2) in a region R, is given by

$$\iint\limits_R \frac{1}{\cos\gamma}\, dx_2\, dx_1,$$

where γ is the angle between the x_3-direction and the normal to the surface ∇f. This result extends to higher dimensions. Therefore, the volume of $\Delta^n(z)$ within the region $0 \leq x_1 \leq y$ is given by

$$V_n(y,z) = \int_0^y \int_0^{z-x_1} \int_0^{z-x_1-x_2} \cdots \int_0^{z-\sum_{i=1}^{n-1} x_i} \frac{1}{\cos\gamma_{n+1}}\, dx_n \cdots dx_3\, dx_2\, dx_1, \tag{3.2}$$

where $0 \leq y \leq z$, and γ_{n+1} is the angle between the unit normal to $\Delta^n(z)$ and the basis vector e_{n+1}. Since the unit normal to $\Delta^n(z)$ is $v = (1/\sqrt{n+1})[1, 1, \ldots, 1]^T$, it follows that

$$\frac{1}{\cos\gamma_{n+1}} = \frac{1}{e_{n+1} \cdot v} = \sqrt{n+1}.$$

The volume $V_n(y,z)$ satisfies the scaling relationship,

$$V_n(y,z) = C^n V_n(y/C, z/C) \tag{3.3}$$

for any constant $C > 0$. This result is obtained by a simple change of variables $\hat{x}_i = x_i/C$ in the integrals in (3.2).

Differentiating (3.2) with respect to y gives

$$\frac{dV_n(y,z)}{dy} = \int_0^{z-y} \int_0^{z-y-x_2} \cdots \int_0^{z-y-\sum_{i=2}^{n-1} x_i} \sqrt{n+1}\, dx_n \cdots dx_3\, dx_2.$$

Now changing the dummy variable names in the integration $x_i \mapsto x_{i-1}$ generates the differential recurrence relation

$$\frac{dV_n(y,z)}{dy} = \sqrt{\frac{n+1}{n}}\, V_{n-1}(z-y, z-y).$$

Applying the scaling relationship (3.3) with $C = (z-y)/z$ to the above gives a simple differential equation for $V_n(y,z)$:

$$\frac{dV_n(y,z)}{dy} = \sqrt{\frac{n+1}{n}} \left(\frac{z-y}{z}\right)^{n-1} V_{n-1}(z,z).$$

Since $V_{n-1}(z,z)$ is constant with respect to y, the above equation can be easily solved along with the initial condition $V_n(0,z) = 0$ giving

$$V_n(y,z) = \sqrt{\frac{n+1}{n}}\, \frac{V_{n-1}(z,z)}{n\, z^{n-1}} \left[z^n - (z-y)^n\right]. \tag{3.4}$$

Noting that $V_1(z,z) = \sqrt{2}z$, substituting $y = z$ in the above gives a recurrence relation for the total volume of $\Delta^n(z)$, which is also easily solved,

$$V_n(z,z) = \sqrt{\frac{n+1}{n}} \frac{V_{n-1}(z,z)\, z}{n} \implies V_n(z,z) = \frac{z^n \sqrt{n+1}}{n!}.$$

Substituting this expression into (3.4) gives the following result for the volume of $\Delta^n(z)$ in the region $0 \leq x_1 \leq y$

$$V_n(y,z) = \frac{\sqrt{n+1}}{n!} \left[z^n - (z-y)^n\right]. \tag{3.5}$$

Therefore, the fraction, ϕ, of the volume of $\Delta^n(z)$ that is in the region $x_1 \geq y$ is given by

$$\phi = 1 - \frac{V_n(y,z)}{V_n(z,z)} = \left(1 - \frac{y}{z}\right)^n.$$

Inverting this relationship gives the value y of x_1 so that the fraction ϕ of the volume of $\Delta^n(z)$ lies to right of x_1:

$$y = \left(1 - \phi^{1/n}\right) z. \tag{3.6}$$

Given the relationship (3.6) it is now possible to describe how to uniformly sample on $\Delta^n(1)$. Define the functions

$$P_i(\phi, z) = \left(1 - \phi^{1/i}\right) z, \qquad 1 \leq i \leq n.$$

Sample ϕ_1 uniformly on $[0,1]$ and set $x_1 = P_n(\phi_1, 1)$. The remaining variables $x_2, \ldots, x_n$ lie in $\Delta^{n-1}(1-x_1)$. So sample ϕ_2 uniformly on $[0,1]$ and set $x_2 = P_{n-1}(\phi_2, 1-x_1)$. Continue in this manner. Thus if the ϕ_i are uniformly sampled on $[0,1]$, define

$$x_i = P_{n+1-i}\left(\phi_i, 1 - \sum_{j=1}^{i-1} x_j\right), \quad 1 \leq i \leq n, \qquad \text{and} \qquad x_{n+1} = 1 - \sum_{j=1}^{n} x_j.$$

Algorithmically the method is defined by

$sum \leftarrow 0$
for i from 1 to n **do**
  $\phi_i \leftarrow$ uniform$[0,1]$
  $x_i \leftarrow (1 - sum)\left(1 - \phi_i^{1/(n+1-i)}\right)$
  $sum \leftarrow sum + x_i$
end for
$x_{n+1} \leftarrow 1 - sum.$

The case where one wishes to uniformly sample on the more general simplex $\sum_{i=1}^{n+1} c_i x_i = z$, where, $z > 0$ and the c_i are positive constants can also be done. The key is to define new variables $\hat{x}_i = x_i c_i / z$, which makes the new variables lie in $\Delta^n(1)$, and then note that since the new variables are linear scalings of the original ones, and since the simplex is flat, that is, has constant normal, the formula for the fraction ϕ of the volume of the simplex to the right of $x_1 = y$ is the same as that for $\hat{x}_1 = \hat{y}$. Thus one simply needs to generate uniform samples $\hat{x}$ on $\Delta^n(1)$ by any of the above three methods and then scale them back via $x_i = \hat{x}_i z / c_i$.

References

[1] J. L. Bentley and J. B. Saxe. Generating sorted lists of random numbers. *ACM Transactions of Mathematical Software*, 6(3):359–364, 1980.

[2] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[3] I. Gerontides and R. L. Smith. Monte Carlo generation of order statistics from general distributions. *Applied Statistics*, 31:238–243, 1982.

[4] D. Lurle and H. O. Hartley. Machine-generation of order statistics for Monte Carlo computations. *The American Statistician*, 26(1):26–27, 1972.

[5] D. Lurle and R. L. Mason. Empirical investigation of several techniques for computer generation of order statistics. *Communications in Statistics*, 2:363–371, 1973.

[6] S. Malmquist. On a property of order statistics from a rectagular distribution. *Skandinavisk Aktuarietidskrift*, 33:214–222, 1950.

[7] M. Rabinowitz and M. L. Berenson. A comparison of various methods of obtaining random order statistics for Monte Carlo computations. *The American Statistician*, 28(1):27–29, 1974.

[8] W. R. Schucany. Order statistics in simulation. *Journal of Statistical Computation and Simulation*, 1(3):281–286, 1972.

[9] P. V. Sukhatme. Tests of significance for samples of the chi square population with two degrees of freedom. *Ann. Eugen.*, 8:52–56, 1937.

Department of Mathematics & Statistics, University of Guelph, Guelph, ON, N1G 2W1, Canada
Email address: AWillms@uoguelph.ca