A Note on the Performance of the "Ahrens Algorithm"

W. Hörmann, Istanbul *

June 3, 2002

Abstract

This short note discusses performance bounds for "Ahrens" algorithm, that can generate random variates from continuous distributions with monotonically decreasing density. This rejection algorithm uses constant hat-functions and constant squeezes over many small intervals. The choice of these intervals is important. Ahrens has demonstrated that the equal area rule that uses strips of constant area leads to a very simple algorithm. We present bounds on the rejection constant of this algorithm depending only on the number of intervals.

AMS Subject Classification: 65C10, 65U05, 11K45
Keywords: non-uniform random variate generation, universal method, performance measures

1 Introduction

It is well known that we can speed up random variate generation by decomposing the density of the desired distribution into many pieces. The simplest way to do this is probably decomposition into many intervals together with rejection from constant hat-functions combined with constant squeezes in every interval. Devroye [3] discusses automatic algorithms based on this main idea under the name "strip methods"; Ahrens [1] calls them "table methods". In [2] Ahrens points out that the algorithm becomes even simpler and especially the tables and the set-up much shorter if we construct the intervals such that the area below the hat remains constant in all intervals. As we like the shortness and simplicity of this algorithm we present a theorem on its performance. As we used the name "Ahrens Algorithm" in our internal discussions for many years we decided to keep it in this paper in honour of Joe Ahrens, one of the great pioneers of random variate generation.

We restrict our attention in the sequel to bounded, monotonically decreasing densities $f(x)$ for $a \leq x \leq b$. If the area below f is not one we call f

*This work was supported by the Austrian Academy of Science, APART-scholarship and by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung,* Proj. 12805-MAT.

a *quasi-density* and write $A_f = \int_a^b f(x)\,dx$. The interval (a,b) will be partitioned into N intervals by the design points $a = p_0 < p_1 < p_2 < \cdots < p_N = b$. Then for the interval (p_{i-1}, p_i) we use the constant hat-function $f(p_{i-1})$ and the constant squeeze $f(p_i)$. Obviously the total area below the hat is given by $A_h = \sum_{i=1}^N f(p_{i-1})(p_i - p_{i-1})$ and the area below the squeeze $A_s = \sum_{i=1}^N f(p_i)(p_i - p_{i-1})$. The expected number of iterations for the algorithm (often called rejection constant) is $\alpha = A_h/A_f$; the expected number of necessary evaluations of f to generate one variate is $E_f = (A_{h-s})/A_f$, where A_{h-s} denotes $A_h - A_s$. For all rejection algorithms we have $\alpha \leq 1 + A_{h-s}/A_f$. Therefore the bounds on A_{h-s} we discuss in this paper can be used to obtain bounds for α and E_f.

Of course the simplest possibility is using equidistant design points. Devroye mainly considers this case in [3, Chapter VIII.2]. As pointed out in [1] it is easy to see that these points are minimizing A_h and A_{h-s} for linear densities. In both references different methods for optimising the design points are discussed as well. As they lead to more complicated algorithms we omit them here and only present the following simple theorem contained in [3]:

Theorem 1. *We are given a monotone quasi-density f on the interval (a,b) with integral not necessarily equal to one. For equidistant partition points we have:*

$$A_{h-s} = \frac{(b-a)(f(a)-f(b))}{N} \tag{1}$$

Proof. Decomposing (a,b) into N intervals of equal length we can easily compute:

$$A_{h-s} = \left(\sum_{i=1}^{N} f\left(\frac{i-1}{N}\right) - f\left(\frac{i}{N}\right)\right)\frac{b-a}{N} = \frac{(f(a)-f(b))(b-a)}{N}$$

□

2 Equal area approach

J. Ahrens suggested to choose the design points such that the area below the hat remains constant for all sub-intervals. In [2] it is dicussed in detail that the automatic algorithm for sampling from a distribution with given density becomes simpler than for equidistant design points for this choice. It is enough to use a single table for storing the p_i as the $f(p_i)$ can be computed from the p_i. In addition it is very simple to choose an interval randomly as the area below the hat is the same for all intervals. So there is no need to use an indexed search or alias method due to the equal area rule.

To compute the construction points p_i using the *equal area rule* the user has to choose c, the constant area below the hat for all sub-intervals first. Then we can easily compute the p_i recursively using

$$p_{i+1} = p_i + c/f(p_i) \qquad \text{and } p_0 = a. \tag{2}$$

We stop with the recursion when the first $p_i \geq b$ occurs. The index of this rightmost design point will be called N. We cannot expect to have $p_N = b$, unless we use a slow search algorithm. So we have to set $f(p_N) = 0$ and have no squeeze in that interval. Note that we must not change p_N as the last interval must have area c below the hat as well.

In [1] it is noted that the equal area approach is minimizing A_h for $f(x) = 1/x$ for arbitrary $0 < a < b < \infty$ (the proof is simple). In [2] a theorem is proven that $A_h - A_f$ can be made arbitrarily small by choosing a large number of intervals, but no comment on the speed of the convergence is given. It is also stated that N will be a bit larger than A_f/c and is demonstrated what this means for serveral examples but the proof is only showing that N is bounded. In the below theorem we answer both questions giving general bounds for the area between hat and squeeze for the equal area approach.

Theorem 2. *We are given a monotone quasi-density (with integral A_f) on the interval (a, b). Using the recursion (2) we construct a piecewise constant hat such that the area below the hat is c in every interval. (To guarantee this we allow $p_N > b$ and set $f(p_N) = 0$). Then we have the following bounds for the area between squeeze and hat A_{h-s} and for the maximal number of intervals N:*

$$A_{h-s} \leq (\log f(a) - \log f(b) + 1)c, \tag{3}$$

$$N \leq \frac{A_f}{c} + \log f(a) - \log f(b) + 1. \tag{4}$$

If $f(b)$ is close or equal to 0 the bounds below (that are also generally correct) will yield better results:

$$A_{h-s} \leq (\log f(a) + \log(b-a) - \log c + 2)c, \tag{5}$$

$$N \leq \frac{A_f}{c} + \log f(a) + \log(b-a) - \log c + 2. \tag{6}$$

Proof. For a fixed quasi-density f and fixed c we can compute for the i-th interval

$$A_{h-s}(i) = (p_i - p_{i-1})(f(p_{i-1}) - f(p_i)) = c\frac{f(p_{i-1}) - f(p_i)}{f(p_{i-1})}$$

using the recursion of the equal area approach defined above. This implies

$$A_{h-s} = c \sum_{i=1}^{N} \left(1 - \frac{f(p_i)}{f(p_{i-1})}\right).$$

Using $1 - x \leq -\log x$ we obtain the bound

$$A_{h-s} \leq c \sum_{i=1}^{N} (\log f(p_{i-1}) - \log f(p_i)) = c(\log f(a) - \log f(p_N))$$

that is valid for all monotone f. The above bound is always correct but p_N need not be equal to b. When we want to replace p_N by b we have to consider that there could be an extra interval to the right of b (which could be fully part of A_{h-s}). We therefore have to add another c thus obtaining the first inequality:

$$A_{h-s} \leq (\log f(a) - \log f(b) + 1)c.$$

The second inequality then follows from $N \leq \frac{A_f + A_{h-s}}{c}$.

If $f(b)$ is zero the above bounds are useless. To get a bound for that case we start from the right border with a first strip of size c. It is clear that the value of f for the left end of that rightmost interval is due to the equal area rule greater or equal $c/(b-a)$. So we can use $f(b) \geq c/(b-a)$ and its reformulation $-\log f(b) \leq \log(b-a) - \log c$ in the first inequality. As we have added one additional tower of size c to the right we have to add one c and thus get:

$$A_{h-s} \leq (\log f(a) + \log(b-a) - \log c + 2)c.$$

As above we obtain the bound for N using $N \leq \frac{A_f + A_{h-s}}{c}$. □

Theorem 2 gives bounds for A_{h-s} and N for the "equal-area" algorithm with arbitrary fixed c. This is the situation a user of that algorithm faces in practice when he has to decide about c and he can use the results directly to get information about the maximal size of arrays needed. It is also possible to fix the maximal N and to compute the corresponding c by interpreting the bounds (4) and (6) as equalities.

If we want to compare the results of the equal area approach with the equidistant points approach we need bounds for A_{h-s} that depend on N. We can obtain them as a simple corollary of Theorem 2. We think that for comparison purposes it is better to assume here that we have chosen c such that the right border of the last sub-interval is equal to b which is possible but not necessary in practice. As formulated in the first section of the proof above this means that we assume that $p_N = b$. Thus we get:

Corollary 3. *We are given a monotone quasi-density (with integral A_f) on the interval (a, b). Using the recursion (2) we construct a piecewise constant hat such that the area below the hat is c in every interval and $p_N = b$. Then we have the following two bounds for the area between squeeze and hat A_{h-s}.*

$$A_{h-s} \leq \frac{A_f}{\frac{N}{k_1} - 1} = A_f \sum_{i=1}^{\infty} \left(\frac{k_1}{N}\right)^i \text{ for } N > k_1,\ k_1 = \log f(a) - \log f(b). \tag{7}$$

$$A_{h-s} \leq \frac{A_f}{\frac{N}{k_2} - 1} = A_f \sum_{i=1}^{\infty} \left(\frac{k_2}{N}\right)^i,\ N > k_2 \tag{8}$$

with $k_2 = \log f(a) + \log(b-a) + \log N - \log A_f + 1$.

Proof. As we assume now that $p_N = b$, we can write the bounds (3) and (5) as $A_{h-s} \leq k\,c$, with $k = \log f(a) - \log f(b)$ or $k = \log f(a) + \log(b-a) + \log N - \log A_f + 1$. In the same way we can write the bounds (4) and (6) as $N \leq \frac{A_f}{c} + k$. So for $N > k$ we get $c \leq \frac{A_f}{N-k}$ and plugging that into the bound for A_{h-s} completes the proof. □

The bound (7) is asymptotically tight if f is continuous. This can be seen from the first part of the proof of Theorem 2 as for continuous densities the ratio $\frac{f(p_i)}{f(p_{i-1})}$ converges to one for all intervals with $c \to 0$ or $N \to \infty$. For densities with discontinuities the bound remains correct but A_{h-s} no longer converges to the upper bound as the ratio $\frac{f(p_i)}{f(p_{i-1})}$ does not converge to one at the points of discontinuity.

That the bound (7) is close to the true values even for moderate N can be seen for $f(x) = 1/x$ in the interval $(1, b)$. For this quasi-density the recursion 2 is so simple that we can find a closed form for $A_{h-s} = N(b^{1/N} + b^{-1/N} - 2)$. It then turns ot that the relative error of the approximation for $N = 100$ is less than 2.5 % for $b = 10$ and less than 5 % for $b = 100$.

It is also clear by our Theorems that for $f(b) > 0$ the rate of convergence for the equal area approach is the same as for equidistant design points. Of course the constants are different and they depend on the distribution. If we consider densities with a long tail and $f(b) > 0$ it is obvious that the equal area approach leads to much better results than intervals of constant length.

This situation changes if $f(b) = 0$ and we get $A_{h-s} = O(\log N/N)$; so for this case the rate of convergence is slower for the equal area approach than for equidistant points. This can be understood by the fact that when using the equal area approach and $f(b) = 0$ the sub-intervals close to b become shorter at a rate slower than $O(1/N)$. For example the last interval becomes shorter at the rate $O(1/\sqrt{N})$ if $f(x)$ is linear and $f(b) = 0$. So it is not astonishing that the equal area approach leads to a slower convergence of A_{h-s} than equidistant design points.

If we want to use our results to compare the performance of the equal area approach with the performance of equidistant design points we should not forget that in practice we are rarely interested in $N > 10^3$; often we just want to fix N eg. to be 100, 500 or 1000 as these allows a simple organisation of the tables in the program. For this case we have to look at the constants of the equation in Theorem 1 and of the bounds in Corollary 3. We understand that for fixed N, A_{h-s} increases with increasing $f(a)$ and increasing $b - a$, ie. if the density gets a sharper peak or a longer tail. To demonstrate that we introduce the family of densities $f(x) = \frac{\gamma+1}{2^{\gamma+1}-1}(2-x)^\gamma$ on $(0, 1)$ for $\gamma \geq 1$. The density is linear for $\gamma = 1$ and gets sharply peaked at 0 for increasing γ; $f(1) = 2/3$ for $\gamma = 1$, and it converges to 0 for increasing γ. Table 2 compares the exact value of A_{h-s} for equidistant design points with the two bounds of Corollary 3 for different values of N and γ. The results show as expected that for the linear density ($\gamma = 1$) equidistant design points are best, but the equal area rule is not much

	$\gamma = 1$	$\gamma = 10$	$\gamma = 10^2$	$\gamma = 10^3$	$\gamma = 10^4$
			$N = 100$		
equidistant exact	0.0067	0.055	0.505	5.005	50.005
equal area bound (7)	0.0070	0.074	2.259	$k > N$	$k > N$
equal area bound (8)	0.0626	0.079	0.105	0.134	0.164
			$N = 1000$		
equidistant exact	0.00067	0.0055	0.0505	0.5005	5.0005
equal area bound (7)	0.00069	0.0070	0.0744	2.2589	$k > N$
equal area bound (8)	0.00826	0.0097	0.0120	0.0143	0.0167

Table 1: A_{h-s} for equidistant design points and the two bounds for A_{h-s} for the equal area rule as given in Corollary 3 for the density $f(x) = \frac{\gamma+1}{2^{\gamma+1}-1}(2-x)^{\gamma}$.

worse indicated by bound (7). For larger values of γ, A_{h-s} of the equal area rule is increasing rapidly, bound (7) is also increasing fast as $f(1)$ approaches 0, only bound (8) increases slowly; it clearly demonstrates that for a density with sharp peak and long tail and moderate fixed N the equal area approach leads to much smaller values for A_{h-s} than equidistant design points. The behaviour of the equal area approach is for such densities similar to the behaviour of the geometrically increasing intervals suggested in [3, Chapter VIII.2.]. Nevertheless we can easily show that for the equal area approach (as for any other choice of the design points) A_{h-s} is not uniformly bounded for N fixed and $\gamma \to \infty$.

3 Conclusions

It is known that the rejection constant is $1 + O(1/N)$ for equidistant points. For the equal area approach we obtained the same result if the density is not zero at b the right endpoint of the interval. If $f(b) = 0$ we get $\alpha = 1 + O(\log N/N)$ for the equal area rule; nevertheless for moderate fixed N and densities with sharp peak or long tail the rejection constant of the equal area approach is much smaller than that for equidistant design points. Considering the simplifications in the sampling algorithm and the reduced storage requirements for the equal area approach our theorems imply that the "Ahrens Algorithm" contains a good strategy to decompose the domain of the distribution.

References

[1] J. H. Ahrens. Sampling from general distributions by suboptimal division of domains. *Grazer Math. Berichte*, 319:20 pp., 1993.

[2] J. H. Ahrens. A one-table method for sampling from continuous and discrete distributions. *Computing*, 54(2):127–146, 1995.

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

Wolfgang Hörmann
Deparment of Industrial Engineering
Bogazici University Istanbul
80815 Bebek-Istanbul
Turkey
hormannw@boun.edu.tr