LOGNORMAL LAW FOR A RENORMALIZATION CHAIN ARISING IN SEARCH THEORY AND THE MODELLING OF DESCENT ALGORITHMS *

J. Michael STEELE

Program in Engineering Statistics, Princeton University, Princeton, NJ 08544, USA

Received May 1986
Revised September 1986

A limit theorem is established for the asymptotic state of a Markov chain arising from an iterative renormalization. The limit theorem is illustrated in applications to the theory of random search and in probabilistic models for descent algorithms. Some special cases are also noted where exact distributional results can be obtained.

Markov chain * lognormal * descent algorithm

1. Introduction

To illustrate the limit theorem which is established here we first consider a simplified question from the theory of random search. We suppose that we seek to find an object which is located at unknown position P. The tools available are assumed to allow us to get closer to P with each probe. The natural goal is to understand how far away from the target we will be with the nth probe.

To specify a model for this process we suppose that $f:[0, 1] \to R$ is a non-negative integrable function and $F(x)$ is the integral of f from 0 to x. If $X_n = x$ is the distance from target of the nth probe, then the distance from the target of the $n+1$st probe is assumed to satisfy

$$P(X_{n+1} \in \mathrm{d}y \mid X_n = x) = f(y)/F(x), \quad 0 < y < x. \tag{1.1}$$

To be still more concrete, we can consider the situation where Z_1 is the radius of a point chosen randomly from the unit disc, Z_2 is a point chosen randomly from the disc of radius $|Z_1|$, and successively Z_n is chosen at random from the disc of radius $|Z_{n-1}|$. To put this in the notation of the more general search model, we can let $F(x) = \pi x^2$ and $f(x) = 2\pi x$, whence it is clear that $|Z_n|$ and X_n have the same probability law.

The random variables $|Z_n|$ can also serve to aid our intuition about the general asymptotic behavior of X_n. An easy calculation shows $\mathrm{E}|Z_n| = (2/3)^n$, and this suggests that X_n is best studied on a logarithmic scale. Still, it is perhaps surprising that for a wide class of functions f, the scaled random variables $Y_n = \log X_n$ are sufficiently tractable to allow a central limit theorem.

Models with the renormalization form given here are quite natural. The random variable X_{n+1} is just a weighted random choice from that mass of f which is left over after the selection of X_n. The defining renormalization relation (1.1) is also closely connected with the basic models of choice theory (see, e.g., Luce [3] and Steele [4]).

Probably the most intriguing application of renormalization chains is to the theory of descent algorithms. That application is the one which is developed here in the most detail.

First, we will establish the main results.

2. Approximate recursion for the characteristic function

Just as the asymptotics of the mean for the simple example on the disc illustrates that X_n should be studied in the log scale, the relationship

Research was supported in part by National Science Foundation Grant no. DMS-8414069.

$f(x) = x$ (or, more generally, $f(x) = x^\alpha$) provides a leading case for the study of a characteristic function associated with X_n. Since our main result contains what might otherwise appear to be a technical hypothesis, we will first derive some formulas which, together with the leading case $f(x) = x^\alpha$, will motivate the final result.

We will let $\phi_n(t)$ be the characteristic function of $\log X_n$ and we will aim toward a recursion relation for the $\{\phi_n\}$. Letting g_n denote the density of X_n, we can calculate

$$\begin{aligned}\phi_n(t) &= \int_0^1 e^{it\log y} P(X_n \in dy)\\ &= \int_0^1 e^{it\log y}\left(\int_y^1 \frac{g_{n-1}(x)}{F(x)}dx\right) f(y)\,dy\\ &= \int_0^1 e^{it\log x} g_{n-1}(x)\\ &\quad\times\left\{\int_0^x \frac{e^{it\log(y/x)}f(y)}{F(x)}dy\right\}dx. \qquad (2.1)\end{aligned}$$

This expression would be a recursion relation if the bracketed integral were independent of x. This will be seen to be approximately the situation. We first explore the inside integral in the leading case $f(x) = x^\alpha$, $\alpha > -1$,

$$\begin{aligned}&\int_0^x \{e^{it\log(y/x)} f(y)/F(x)\}\,dy\\ &= (\alpha+1)\int_0^x e^{it\log(y/x)} y^\alpha x^{-\alpha-1}\,dy\\ &= (1+\alpha)\int_0^1 e^{it\log u} u^\alpha du\\ &= \left(1+\frac{it}{1+\alpha}\right)^{-1}. \qquad (2.2)\end{aligned}$$

In the motivating case of $|Z_n|$, we saw $\alpha = 1$; so, the recursion (2.1) and the identity (2.2) give us that the characteristic function of $\log|Z_n|$ is exactly equal to $(1 + it/2)^{-n}$, *i.e.*, $-\log|Z_n|$ is equal to the sum of n exponential random variables with mean $\frac{1}{2}$. Naturally, this result can be obtained more directly, but the explicit example provides a conceptual guide and analytic check on our main calculation. As a final note on the example, we see that $\log E|Z_n| = n\log(2/3)$ is consistent with $E\log|Z_n| = -n/2$ since Jensen's inequality states $E\log|Z_n| \le \log E|Z_n|$ and indeed $-0.5 \le \log(2/3) = -0.405$.

We now let $\psi(t) = (1 + it/(1+\alpha))^{-1}$ and combine eqs. (2.1) and (2.2) to obtain

$$\begin{aligned}&\phi_n(t) - \psi(t)\phi_{n-1}(t)\\ &\quad= \int_0^1 e^{it\log x} g_{n-1}(x) h(x)dx, \qquad (2.3a)\end{aligned}$$

where

$$h(x) = \int_0^x e^{it\log(y/x)}\left\{\frac{f(y)}{F(x)} - \frac{y^\alpha(\alpha+1)}{x^{\alpha+1}}\right\}dy. \qquad (2.3b)$$

We can now use (2.3) to get an effective measure of the extent to which $f(x)$ differes from x^α. We have

$$\begin{aligned}&\left|\int_0^1 e^{it\log x} g_n(x)h(x)dx\right|\\ &\quad\le \int_0^1 g_n(x)\,|h(x)|\,dx\\ &\quad\le \int_0^1 g_n(x)h^*(x)\,dx = r_n, \qquad (2.4a)\end{aligned}$$

where we define

$$h^*(x) = \int_0^x \left|\frac{f(y)}{F(x)} - \frac{(\alpha+1)y^\alpha}{x^{\alpha+1}}\right| dy. \qquad (2.4b)$$

The quantity h^* captures just what we need to know about the approximability of $f(x)$ by x^α. We now use the defining recursion for g_n to calculate a recursion for r_n as follows:

$$\begin{aligned}r_n &= \int_0^1 g_n(x)h^*(x)dx\\ &= \int_0^1\left\{\int_x^1 g_{n-1}(z)\frac{f(x)}{F(z)}dz\right\}h^*(x)dx\\ &= \int_0^1 g_{n-1}(z)\left\{\frac{1}{F(z)}\int_0^z f(x)h^*(x)dx\right\}dz. \qquad (2.5)\end{aligned}$$

So, if there is a $0 < \lambda < 1$ such that

$$\frac{1}{F(z)}\int_0^z f(x)h^*(x)dx \le \lambda h^*(z) \qquad (2.6)$$

for all $z \in [0, 1]$, then

$$r_n \le \lambda r_{n-1} \quad\text{and}\quad r_n \le \lambda^n r_0. \qquad (2.7)$$

The recursion relations (2.3a) and (2.7) can be used to approximate ϕ_n in terms of powers of ψ by noting

$$\begin{aligned}&|\phi_n(t) - \psi(t)\phi_{n-1}(t)| \le r_{n-1} \le \lambda^{n-1} r_0,\\ &|\psi(t)\phi_{n-1}(t) - \psi^2(t)\phi_{n-2}(t)| \le r_{n-2} \le \lambda^{n-2} r_0,\end{aligned}$$

and generally,

$$\left|\psi^{m-1}(t)\phi_{n-m+1}(t)-\psi^{m}(t)\phi_{n-m}(t)\right| \le r_{n-m}\le \lambda^{n-m}r_0.$$

Summing the expressions above for m between 1 and $n-k$, we have

$$\left|\phi_n(t)-\psi^{n}(t)\phi_k(t)\psi^{-k}(t)\right| \le r_0\lambda^k(1-\lambda)^{-1}. \tag{2.8}$$

Here one should note that the characteristic function $\psi(t)$ does not vanish for any real t, $\psi'(0)=-i(1+\alpha)^{-1}$, and $\psi''(0)=-2(1+\alpha)^{-2}$. If W is an exponential random variable with mean μ equal to $(1+\alpha)^{-1}$ and variance σ^2 equal to $(1+\alpha)^{-2}$ then ψ is the characteristic function of $-W$. We see therefore that $\psi(t(1+\alpha))e^{it}$ is the characteristic function of a random variable with mean zero and variance 1. By the central limit theorem, we then get that $\psi^n(t(1+\alpha)/n^{1/2})e^{it\sqrt{n}}$ converges to $e^{-t^2/2}$ for all t, so (2.8) implies for all fixed k that

$$\limsup_{n\to\infty}\left|\phi_n\left(\frac{t(1+\alpha)}{\sqrt{n}}\right)e^{it\sqrt{n}}-e^{-t^2/2}\right| \le r_0\lambda^k(1-\lambda)^{-1}.$$

Since k is arbitrary, the required limit theorem is established. We can now summarize the main result.

Theorem. *If $f:[0,1]\to R$ is a non-negative function and $F(x)=\int_0^x f(y)\mathrm{d}y<\infty$, $0\le x\le 1$, then for the Markov chain $\{X_n\}_{n=0}^{\infty}$ defined by the kernel*

$$K(x,y)=P(X_{n+1}\in \mathrm{d}y \mid X_n=x)=f(y)\mid F(x),\quad 0<y<x,$$

we have the convergence in distribution of

$$\frac{\log X_n+n/(1+\alpha)}{\sqrt{n}/(1+\alpha)} \tag{2.9}$$

to the standard normal, provided only that the function

$$h^*(x)=\int_0^x\left|\frac{f(y)}{F(x)}-\frac{(\alpha+1)y^{\alpha}}{x^{\alpha+1}}\right|\mathrm{d}y$$

satisfies the integral inequality

$$\frac{1}{F(z)}\int_0^z f(x)h^*(x)\,\mathrm{d}x<h^*(z) \tag{2.10}$$

for all $0<z\le 1$.

Here one should note that there is nothing special about the interval [0, 1]. It was chosen just to keep notation clean. Also, the continuity of h^* guarantees that if (2.10) holds, then the nominally stronger uniform inequality (2.6) must hold as well. The effective use of the preceding result will often require a proper choice of scale in order to be able to establish (2.10), and an instance of such change of scale is given in the course of the examples detailed in the next section.

3. Random descent algorithms

If ϕ is a convex function on R^d, then a sequence of vectors $x_n\in R^d$, $1\le n<\infty$, is said to satisfy the descent condition with respect to ϕ provided $\phi(x_{n+1})<\phi(x_n)$. Many important algorithms are known to produce sequences which satisfy a descent condition for an appropriate choice of ϕ, and the descent property is one of the basic tools for analyzing numerical algorithms (cf. Luenberger [2]). The EM algorithm is an example of such a descent algorithm currently of great interest in statistics (see, e.g., Dempster, Rubin and Laird [1] and Wu [5]). In addition to calling attention to the presence of the descent condition, our reason for singling out the EM algorithm is that it is known to exhibit linear convergence, as opposed to the quadratic convergence which is typical of Newton method based algorithms. The lognormal law which was established in the preceding section will now be used to show that, in general, random descent algorithms have a very specific type of linear convergence.

Here, for specificity, we will assume that $\phi\ge 0$ and that 0 is the minimum value attained by ϕ in the convex set $\{x:\phi(x)\le 1\}$. We can then define a random descent sequence by choosing a point U_1 at random from $\{x:\phi(x)\le 1\}$, and then successively choosing U_{n+1} at random from $\{x:\phi(x)\le\phi(U_n)\}$.

If we let μ denote Lebesque measure on R^d, then we will shortly show that the main theorem of this note establishes that $X_n=\mu\{x:\phi(x)\le\phi(U_n)\}$ is asymptotically lognormal for a large class of convex functions ϕ.

We first examine the special case of quadratic ϕ. In this instance the expression $-\log X_n$ can be shown to have *exactly* a gamma distribution.

If Q is a symmetric positive definite matrix, the

quadratic form x^TQx defines a convex function ϕ which can also be written as $\phi(x) = x^TM^TMx$ where M is a non-singular linear mapping of R^d onto R^d. Setting $F(t) = \mu\{x : \phi(x) \le t\}$, we see

$$F(t) = \mu M^{-1}\{x : x^Tx \le t\} = \omega_d t^d (\det M)^{-1} = \omega_d t^d (\det Q)^{-1/2}, \tag{3.1}$$

where ω_d is the volume of the unit sphere in R^d. By the characteristic function calculation given by eq. (2.2), we see that $-\log X_n$ is the sum of n independent exponentials with mean $1/d$.

We will now pursue the general case and build on the fact that any regular convex function is locally well approximated as a quadratic. Here, the notion of approximation needs to be developed with an eye toward the condition (2.10); but, in sympathy with the leading case (3.1), we will first consider those convex ϕ which as $t \to 0$ satisfy the condition

$$f(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mu\{x : \phi(x) \le t\} = ct^{d-1} + 0(t^{d-1+\epsilon}), \tag{3.2}$$

where c and $\epsilon > 0$ are constants. In applications where ϕ is more general than the quadratic leading to (3.1), one can still often obtain (3.2), even with $\epsilon = 1$. The key issue is that of relating condition (3.2) with the more technical hypothesis (2.10) of the main theorem.

From (3.2) we have $F(t) = ct^d/d + 0(t^{d+\epsilon})$, so setting $\alpha = d - 1$, we note that for $0 < y < x$

$$\frac{f(y)}{F(x)} - \frac{(\alpha+1)y^\alpha}{x^{\alpha+1}} = 0\left(y^{\alpha+\epsilon}/x^{\alpha+1}\right), \tag{3.3}$$

and for $h^*(x)$ defined as in (2.4b) we have $h^*(x) = 0(x^\epsilon)$.

If we let $f_\delta(x) = f(\delta x)$, then the corresponding expressions for F_δ and h_δ^* satisfy $F_\delta(x) = \int_0^x f_\delta(u)\mathrm{d}u = \delta^{-1}F(\delta x)$ and

$$h_\delta^*(x) = \int_0^x \left|\frac{f_\delta(y)}{F_\delta(x)} - \frac{(\alpha+1)y^\alpha}{x^{\alpha+1}}\right|\mathrm{d}y = \int_0^{\delta x}\left|\frac{f(u)}{F(\delta x)} - \frac{(\alpha+1)u^\alpha}{(\delta x)^{\alpha+1}}\right|\mathrm{d}u = h^*(\delta x).$$

The basic convergence hypothesis (2.10) applied to f_δ is therefore equivalent to the condition that the integral inequality

$$\int_0^z f(u)h^*(u)\mathrm{d}u < h^*(z) \tag{3.4}$$

hold for all z in the restricted range $(0, \delta]$.

The fact that $h^*(x) = 0(x^\epsilon)$ and the identity (3.2) let us check easily that (3.4) holds for a sufficiently small $\delta > 0$. By the equivalences just established, f_δ and h_δ^* satisfy eq. (2.10).

If f_δ is associated with the Markov chain Y_n by the basic relation (1.1), then Y_n has the same distribution transition kernel as $\delta^{-1}X_n$ for $X_n \le \delta$. Since Y_n is asymptotically lognormal and since $P(X_n > \delta)$ tends to zero, we see X_n is also asymptotically lognormal.

To summarize, we have proved that if μ and ϕ satisfy (3.2), then the random variables defined by

$$Z_n = -\log \mu\{x : \phi(x) \le \phi(U_n)\}$$

have approximate mean n/d and asymptotic variance n/d^2. Moreover, the standardized variables $(Z_n - n/d)/\sqrt{n/d^2}$ are asymptotically normal.

4. Concluding remark

We have exhibited a lognormal law for the asymptotic state of a class of Markov chains which have a transition kernel defined by a natural renormalization. The Markov chain is motivated by a simple search model and it also serves to model the behavior of random descent algorithms. These random descent models exhibit linear convergence of a very precise type which is made explicit by the lognormal law. It would be of interest to provide a random descent model which could exhibit the quadratic convergence rates typical of Newton methods.

References

[1] A.P. Dempster, D.B. Rubin and N.M. Laird, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Statistical Society B* **39**, 1–22 (1977).

[2] D.G. Luenberger, *Linear and Nonlinear Programming*, 2nd ed., Addison-Wesley, Reading, MA, 1984.

[3] R.D. Luce, *Individual Choice Behavior: A Theoretical Analysis*, Wiley, New York, 1959.

[4] J.M. Steele, "Limit properties of Luce's choice theory", *J. Math. Psych.* **11** (2), 124–131 (1974).

[5] C.F.J. Wu, "On the convergence properties of the EM algorithm", *Annals of Statistics* **11**, 95–103 (1983).