91

March 11, 1988
To Appear in: Stochatic
Models

A PROBABILISTIC MODEL FOR THE TIME TO UNRAVEL A STRAND OF DNA

R. J. Vanderbei
L. A. Shepp

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

A common model for the time σ_L (sec) taken by a DNA strand of length L (cm) to unravel is to assume that new points of unraveling occur along the strand as a Poisson process of rate λ $1/(\text{cm}\times\text{sec})$ in space-time and that the unraveling propogates at speed $v/2$ (cm/sec) in each direction until time σ_L.

We solve the open problem to determine the distribution of σ_L by finding its Laplace transform and using it to show that as $x = L^2\lambda/v \to \infty$, σ_L is nearly a constant,

$$\sigma_L = \left[\frac{1}{\lambda v}\log(\frac{L^2\lambda}{v})\right]^{1/2}.$$

We also derive (modulo some small gaps) the more precise limiting asymptotic formula: for $-\infty < \theta < \infty$

$$P\left[\sigma_L < \frac{1}{\sqrt{\lambda v}}\left[\psi^{1/2}(\log(L^2\lambda/v)) + \frac{\theta}{\psi^{1/2}(\log(L^2\lambda/v))}\right]\right] \to e^{-e^{-\theta}},$$

where ψ is defined by the equation:

$$\psi(x) = \log\psi(x) + x, \qquad x \geq 1.$$

These results are obtained by interchanging the role of space and time to uncover an underlying Markov process which can be studied in detail.

1. INTRODUCTION

The problem we are interested in is to find a relation between the length of a strand of DNA and the time it takes for such a strand to unravel. Unraveling is triggered by the release of an enzyme into the surrounding cytoplasm. According to the usual model for unraveling of DNA, we assume that there are a large number of enzymes released and they come into contact with the strand of DNA at random places and at random times. If an enzyme contacts the strand at a place which has yet to unravel, it begins unraveling in both directions from the point of contact at a uniform rate.

We assume that the strand has length L and that there is a Poisson point process with intensity λ in space-time, $[0,L]\times[0,\infty)$, representing the appearance of an enzyme at a particular place and time. Choosing convenient units, we may assume that each point of unraveling propogates at a rate of $1/2$.

Throughout most of the paper we make an important simplifying assumption. Namely, we assume that at time zero the strand begins unraveling at both its endpoints. This is an unrealistic assumption but it makes the mathematics easier. In Section 4, we will show that our asymptotic results for this altered model agree with those for the real model.

A schematic picture of the process is shown in Figure 1. The Poisson points are shown as stars. Note that some points are irrelevant since the DNA has already unraveled at the specified point by the time the enzyme arrived. Nonetheless, in this picture there are four arrivals that do cause additional unraveling. The time at which the entire chain is unraveled is represented by the highest point on the jagged curve.

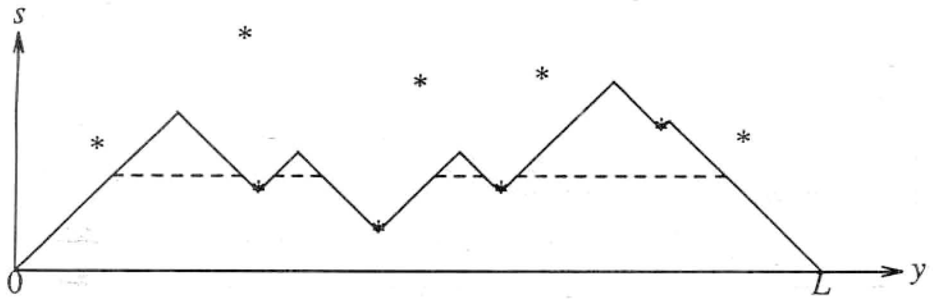

Figure 1. Raveled DNA vs Time

The problem is to find the distribution of the height of the jagged curve as a function of the length L of the strand of DNA.

Before we can begin our analysis, we must make a simple change of coordinates. In words, what we need to do is imagine that we are drifting through the cytoplasm at a rate equal to the rate at which one end unravels and in a direction parallel to the strand. This amounts to making a shear transformation in space-time. That is, if we use y for the spatial coordinate and s for the temporal coordinate, then the transformation is $y' = y + s/2, \quad s' = s$. For notational convenience, we drop the primes in the new coordinate system. The new picture for the problem is shown in Figure 2.

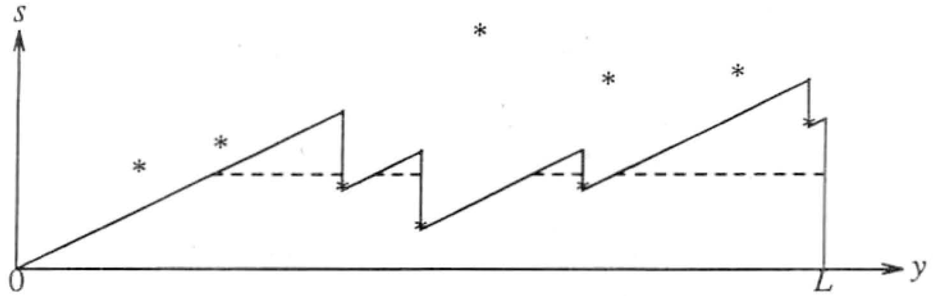

Figure 2. After Shearing

The Poisson point process is invariant under our shear transformation (and, in fact, any measure preserving transformation). Also, now the strand appears to be unraveling only at one end and at a unit rate. A moment's thought reveals that the number and length of the intervals remaining at any time s are the same in both Figures 1 and 2.

The important observation to make is that the jagged line that we have been drawing is now a trajectory of a Markov process. The Markov process X_t can be described as follows. Basically, it consists of uniform upward motion at unit speed with occasional jumps downward. The downward jumps are controlled by an exponential clock. The exponential clock ticks at a rate proportional to the height of the process. In fact the jump rate is precisely λX_t. The new position is uniformly distributed on the interval $[0,X_t]$. Note that the time scale of the Markov process is the space scale for the actual DNA problem. This can potentially cause some confusion. So whenever we refer to "time" for X_t we will use quotes to emphasize that it is not time in the original problem.

For such a simple process, it is easy to write down the transition semigroup:

$$T_t f(x) = (1-\lambda x t) f(x+t) + \lambda x t \int_0^x f(y) \frac{dy}{x} + o(t). \tag{1.1}$$

From (1.1), we see that the infinitesimal generator is given by

$$Af(x) = f'(x) + \lambda \int_0^x (f(y) - f(x)) dy. \tag{1.2}$$

As an aside, we note that it is easy to compute the stationary distribution for this Markov process. Indeed, letting X_∞ denote a random variable with this distribution, the limiting distribution turns out to be semi-Gaussian:

$$P\{X_\infty \geq x\} = e^{-\lambda x^2/2}. \tag{1.3}$$

We return now to the matter at hand. Put

$$\sigma_L = \sup_{0< t \leq L} X_t. \tag{1.4}$$

Our main result is the following.

THEOREM. *As* $L \to \infty$, $E\sigma_L^n \sim (\lambda^{-1}\log L^2\lambda)^{n/2}$.

This theorem is proved in Section 2.

Note that σ_L is almost not random for large L. More precisely, in Section 3, we derive (modulo some small gaps) the following limiting asymptotic formula for σ_L:

$$P\left[\sigma_L < \frac{1}{\sqrt{\lambda}}\left[\psi^{1/2}(x) + \frac{\theta}{\psi^{1/2}(x)}\right]\right] \to e^{-e^{-\theta}}, \tag{1.5}$$

where $x = \log(L^2\lambda)$, $-\infty < \theta < \infty$ and ψ is the unique function characterized as the larger of the two solutions to

$$\psi(x) = \log \psi(x) + x, \qquad x \geq 1.$$

We remark that the right side of (1.5) is similar to the usual maximum laws that appear in the theory of maxima for stationary Gaussian processes [1]. This is perhaps not surprising when one realizes that σ_L is defined as the maximum of a Markov process that, while not quite Gaussian, is asymptotically semi-Gaussian. Note that (1.5) shows that σ_L is for large L nearly the constant $\frac{1}{\sqrt{\lambda}}\psi^{1/2}(x)$ with a random error that tends to zero like $1/\psi(x)$.

Finally, we note that the convergence rate in our main theorem and in (1.5) could be slow and that this could have an impact on applying this formula to real problems. We leave the study of the rate of convergence to future endeavors.

2. *MOMENT CALCULATION*

This section is devoted to the proof of our main theorem. First, we introduce, for each $s \geq 0$, the first hitting time of the level s:

$$\tau_s = \inf \{t : X_t \geq s\}. \tag{2.1}$$

There is a natural duality between σ_L and τ_s:

$$\{\sigma_L \geq s\} = \{\tau_s \leq L\}. \tag{2.2}$$

The advantage of introducing τ_s is that we can explicitly evaluate its Laplace transform. Indeed, for each fixed $\alpha > 0$, let

$$f(x) = E_x e^{-\alpha\tau_s},$$

(we use E_x for expectations calculated when the process starts at x and we use E to denote expectation computed when the process X_t starts at zero). It is well known that f is the solution to the following equations:

$$Af(x) = \alpha f(x), \qquad 0 \leq x \leq s \tag{2.3}$$

$$f(s) = 1. \tag{2.4}$$

Since the operator A is not purely a differential operator, it is desirable to differentiate (2.3) once:

$$f''(x) - \lambda x f'(x) = \alpha f'(x), \qquad 0 \le x \le s. \tag{2.5}$$

We need to introduce one extra condition to guarantee that a solution to (2.5), (2.4) is also a solution to (2.3), (2.4). The extra condition is

$$f'(0) = \alpha f(0). \tag{2.6}$$

It is easy to solve the system (2.5), (2.4), (2.6). The solution is:

$$f(x) = \frac{1 + \alpha \int\limits_0^x e^{\alpha u + \lambda u^2/2} du}{1 + \alpha \int\limits_0^s e^{\alpha u + \lambda u^2/2} du}.$$

Of course, we are mostly interested in $x = 0$:

$$Ee^{-\alpha\tau_s} = f(0) = \frac{1}{1 + \alpha \int\limits_0^s e^{\alpha u + \lambda u^2/2} du}. \tag{2.7}$$

We can use (2.2) to write (2.7) in terms of σ_t:

$$\begin{aligned} Ee^{-\alpha\tau_s} &= \int\limits_0^\infty e^{-\alpha t} \frac{d}{dt} P\{\tau_s \le t\} dt \\ &= \int\limits_0^\infty e^{-\alpha t} \frac{d}{dt} P\{\sigma_t \ge s\} dt. \end{aligned} \tag{2.8}$$

Equating the right-hand side in (2.7) to the right-hand side in (2.8), multiplying both sides by ns^{n-1} and integrating, we get

$$\int\limits_0^\infty e^{-\alpha t} \frac{d}{dt} E\sigma_t^n dt = \int\limits_0^\infty \frac{ns^{n-1} ds}{1 + \alpha \int\limits_0^s e^{\alpha u + \lambda u^2/2} du}. \tag{2.9}$$

Let $I_n(\alpha)$ denote the integral on the right in (2.9). We can now use standard Tauberian arguments to uncover the asymptotic behavior of $E\sigma_t^n$ as t tends to ∞ by studying the behavior of $I_n(\alpha)$ as α tends to 0.

Let

$$\beta(\alpha) = \left[\frac{1}{\lambda}\log\frac{\lambda}{\alpha^2}\right]^{½}.$$

LEMMA 1. (Upper bound)

$$\limsup_{\alpha\to 0} \; I_n(\alpha)/\beta^n(\alpha) \le 1.$$

Proof. For s near 0, a good lower bound on the denominator of the integrand defining $I_n(\alpha)$ is simply 1. For large values of s, we need a more careful estimate. Clearly,

$$\int_0^s e^{\alpha u + \lambda u^2/2} du \ge \int_0^s e^{\lambda u^2/2} du$$

$$\ge e^{\lambda(s-1)^2/2},$$

(the second inequality holding whenever $s \ge 1$). Hence, we can split the range of integration into two parts: one from 0 to $\beta + 1$ and the other from $\beta + 1$ to ∞ (where β is some fixed positive quantity). Doing this, we get

$$I_n(\alpha) \le \int_0^{\beta+1} ns^{n-1} ds + \frac{1}{\alpha}\int_{\beta+1}^{\infty} ns^{n-1} e^{-\lambda(s-1)^2/2} ds.$$

Since $(s-1)/s \ge \beta/(\beta+1)$ for $s \ge \beta + 1$, we see that

$$I_n(\alpha) \le (\beta+1)^n + \frac{n}{\alpha}(\frac{\beta+1}{\beta})^{n-1} \int_{\beta+1}^{\infty} (s-1)^{n-1} e^{-\lambda(s-1)^2/2} ds$$

$$= (\beta+1)^n + \frac{n}{\alpha}(\frac{\beta+1}{\beta})^{n-1} \frac{1}{\lambda^{(n-1)/2}} \int_{\sqrt{\lambda}\beta}^{\infty} r^{n-1} e^{-r^2/2} ds.$$

Now all we need to do is estimate the tail of the n^{th} moment of a normal distribution:

$$\phi_n(x) = \int_x^\infty r^n e^{-r^2/2} dr.$$

We use the usual estimate (see, e.g., [2] Vol 1, p. 175):

$$\phi_n(x) = x^{n-1} e^{-x^2/2} O(1).$$

Applying this to the situation at hand, we get

$$I_n(\alpha) \leq (\beta + 1)^n + \frac{n}{\alpha}(1 + \frac{1}{\beta})^{n-1} \frac{1}{\lambda^{(n-1)/2}} (\sqrt{\lambda}\beta)^{n-2} e^{-\lambda\beta^2/2} O(1).$$

Now, picking $\beta = \beta(\alpha)$, we see that

$$I_n(\alpha) \leq (1 + \frac{1}{\beta(\alpha)})^n + n(1 + \frac{1}{\beta(\alpha)})^{n-1} \frac{O(1)}{\log \frac{\lambda}{\alpha^2}}.$$

Since $\lim_{\alpha \to 0} \beta(\alpha) = \infty$, we finally get that

$$\limsup_{\alpha \to 0} I_n(\alpha)/\beta^n(\alpha) \leq 1.$$

The lower bound is easier to obtain.

LEMMA 2. (Lower bound)

$$\liminf_{\alpha \to 0} I_n(\alpha)/\beta^n(\alpha) \geq 1.$$

Proof. Making crude estimates, we see that

$$I_n(\alpha) \geq \int_0^{\beta} \frac{ns^{n-1}ds}{1 + \alpha \int_0^{s} e^{\alpha u + \lambda u^2/2} du}$$

$$\geq \frac{\beta^n}{1 + \alpha\beta e^{\alpha\beta + \lambda\beta^2/2}}.$$

Now fix $\varepsilon > 0$ and put

$$\beta_\varepsilon(\alpha) = \sqrt{1-\varepsilon}\;\beta(\alpha) = \left(\frac{1-\varepsilon}{\lambda}\log\frac{\lambda}{\alpha^2}\right)^{1/2}.$$

If we pick $\beta = \beta_\varepsilon(\alpha)$, we get

$$I_n(\alpha) \geq \frac{\beta_\varepsilon^n(\alpha)}{1 + \alpha^\varepsilon \beta_\varepsilon(\alpha) e^{\alpha\beta_\varepsilon(\alpha)} \lambda^{(1-\varepsilon)/2}}.$$

Since $\lim_{\alpha\to 0} \alpha^\delta \beta_\varepsilon(\alpha) = 0$ for any $\delta > 0$, we see that

$$1 \leq \liminf_{\alpha\to 0} \frac{I_n(\alpha)}{\beta_\varepsilon^n(\alpha)} = \frac{1}{(1-\varepsilon)^{n/2}} \liminf_{\alpha\to 0} \frac{I_n(\alpha)}{\beta^n(\alpha)}.$$

Since ε is arbitrary, this completes the proof of Lemma 2.

It follows from Lemmas 1 and 2 that

$$\lim_{\alpha\to 0} \frac{I_n(\alpha)}{\beta^n(\alpha)} = 1.$$

Since $\beta(\alpha)$ is slowly varying at zero, the Tauberian theorem implies that (see, e.g. [2] Vol. 2, p. 445)

$$\lim_{t\to\infty} \frac{E\sigma_t^n}{\beta^n(1/t)} = 1.$$

This completes the proof of our main theorem.

3. ASYMPTOTIC LIMITING DISTRIBUTION

Here we give a derivation (with some small gaps) of the actual limiting asymptotics of the time to unravel. The formula we are after is given in (1.5).

Fix $s \geq 0$ and put

$$p(t) = \begin{cases} P\{\tau_s > t\} = P\{\sigma_t < s\} & t \geq 0 \\ 0 & t < 0. \end{cases}$$

Then

$$\int_{-\infty}^{\infty} e^{-\alpha t} p(t) dt = \int_0^{\infty} e^{-\alpha t} P\{\tau_s > t\} dt$$

$$= \frac{1}{\alpha}(1 - Ee^{-\alpha\tau_s})$$

$$= \frac{\int_0^s e^{\alpha u + \lambda u^2/2} du}{1 + \alpha \int_0^s e^{\alpha u + \lambda u^2/2} du} \tag{3.1}$$

Equation (3.1) holds not only for real numbers $\alpha \geq 0$, but also for complex numbers for which $\mathrm{Re}(\alpha) \geq 0$. In fact, we can make the substitution $\beta = i\alpha$ to rewrite (3.1) as a Fourier transform of p. Inverting the Fourier transform we get

$$p(t) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{zt} \frac{\int_0^s e^{zu + \lambda u^2/2} du}{1 + z\int_0^s e^{zu + \lambda u^2/2} du} dz.$$

The integral may be written as a sum of the residues extracted as the contour along the imaginary axis is shifted left. The main contribution is at the first zero of the denominator:

$$f(z) = 1 + z\int_0^s e^{zu+\lambda u^2/2}du.$$

Letting a denote this *principal zero* of f, we see immediately that

$$\int_0^s e^{au+\lambda u^2/2}du = -\frac{1}{a}. \tag{3.2}$$

Using (3.2), it is easy to see that

$$f'(a) = \frac{a}{\lambda}e^{as+\lambda s^2/2} - \frac{1}{a}.$$

The residue at a is

$$\mathrm{Res}(a) = -\frac{e^{at}}{af'(a)}. \tag{3.3}$$

Our goal is to derive the asymptotic formula (1.5). Hence, we will eventually let s be a function of t that tends to infinity as t tends to infinity. Once s becomes a function of t, the principal zero a also becomes a function of t. We should think of t as very large, and s as roughly $(\lambda^{-1}\log t^2\lambda)^{1/2}$; smaller than t but larger than those variables not running off to infinity.

With these relative sizes in mind, let's investigate the principal zero a. Our first observation is that the principal zero is a negative real. To see this, note that from the definition of $f(z)$ and formula (3.1) we have

$$f(z) = \frac{1}{Ee^{-z\tau_s}}.$$

Hence, for any $z = x + iy$ lying in the negative half plane, but not on the real axis,

$$|f(z)| = \frac{1}{|Ee^{-z\tau_s}|} > \frac{1}{Ee^{-x\tau_s}} = f(x).$$

Therefore, for z to be a root, $f(x)$ must be strictly negative. Since $f(0) = 1$, it follows that the principal root must lie on the negative real axis.

The next order of business is to decide whether a approaches zero or infinity as t tends to infinity and then to get a good approximation. We have already noted that $f(0) = 1$. In addition,

$$f'(0) = \int_0^s e^{\lambda u^2/2} du.$$

Recalling that s will be tending to infinity, we see that the slope of f at zero is very large. Hence, we would expect a to be close to zero and approaching zero as t tends to infinity. Consequently, we can use one step of Newton's method to get a reasonable estimate for a:

$$a \approx -\frac{1}{f'(0)} \approx -\lambda s e^{-\lambda s^2/2}. \qquad (3.4)$$

Using this approximation for a, we see that

$$af'(a) \approx -1 + \lambda s^2 e^{-\lambda s^2/2 - \lambda s^2 e^{-\lambda s^2/2}} \approx -1. \qquad (3.5)$$

Substituting (3.4) and (3.5) into (3.3), we get

$$P\{\sigma_t < s\} \approx e^{-\lambda s t e^{-\lambda s^2/2}}. \qquad (3.6)$$

Equating the right side of (3.6) to $e^{-e^{-\theta}}$ we see that s must satisfy

$$\lambda s^2/2 = \theta + \log \lambda s t$$

It is more convenient to rewrite this in the following equivalent manner:

$$\lambda s^2 = 2\theta + \log \lambda s^2 + \log \lambda t^2.$$

From the definition of ψ, we see that

$$\lambda s^2 = \psi(2\theta + \log \lambda t^2).$$

Since 2θ is small potatoes compared to $\log \lambda t^2$, we make the following approximation:

$$\lambda s^2 \approx \psi(x) + \psi'(x) 2\theta,$$

where $x = \lambda t^2$. It follows from the definition of ψ that

$$\psi'(x) = \frac{\psi(x)}{\psi(x) - 1}.$$

Since $\psi(x)$ tends to infinity as x does, it follows that, for large x,

$$\psi'(x) \approx 1.$$

Hence,

$$\lambda s^2 \approx \psi(x) + 2\theta,$$

and so,

$$s \approx \frac{1}{\sqrt{\lambda}} \left\{ \sqrt{\psi(x)} + \frac{\theta}{\sqrt{\psi(x)}} \right\}.$$

This completes the derivation of (1.5).

Note that (1.5) is consistent with the moment theorem of Section 1. This follows from the fact that $\psi(x) \sim x$.

We end this section by reiterating that there are certain gaps that must be filled before formula (1.5) for the asymptotic limiting distribution is rigorous. The most important gap is that the principal residue does indeed dominate. The other gaps are mostly that the various approximations are sufficiently precise. We leave it to future research to fill in these details.

4. *LOOSE ENDS*

We have assumed that unraveling starts from each end at $s = 0$. But in the real model for DNA, this is not the case. However, we show here that the probability that an endpoint is not unraveled by time $T = E\sigma_L$ is asymptotically negligible so that there is no change needed in the formulas given above for the limiting moments or the limiting distribution of σ_L in the real model.

A schematic of the real model is shown in Figure 3. The left endpoint is not unraveled at time T if and only if there are no stars in the triangle bounded by the line connecting the point T on the vertical axis with the point $T/2$ on the horizontal axis. The probability of this event is

$$\pi = P(\text{left endpoint is not unraveled at time T}) = e^{-\lambda T^2/4}.$$

Now, if we substitute the asymptotic formula for $E\sigma_L$ in for T, we get

$$\pi = \frac{1}{(L^2\lambda)^{1/4}}.$$

Clearly, π tends to zero as L tends to infinity. The right endpoint behaves the same.

It is also interesting to determine the expected number of unraveling segments at each time s. It is easiest to work with the sheared model shown in Figure 2. The number N_s of unraveling segments at time s is equal to the number of downcrossings of the level s if X_L is below s and it is one more than that if X_L is above s. Since we don't really mind being off by one, let's simply try to find the expected number of downcrossings of the level s by the "time" L.

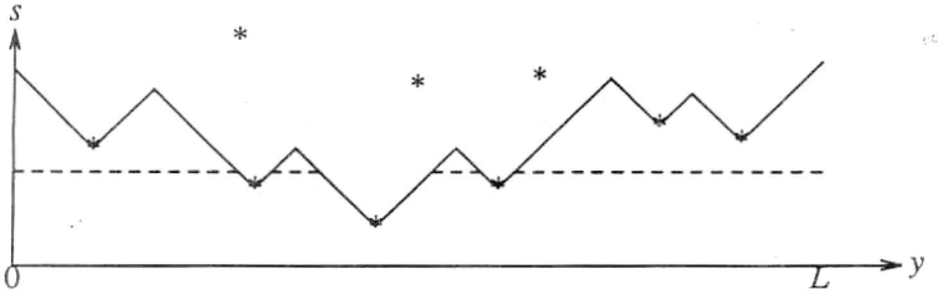

Figure 3. The Real Model

A downcrossing occurs within a small increment of "time" dy if and only if there is a Poisson point in the infinitesimal rectangle $dy \times [0, s]$, which has probability $\lambda s dy$, and there are no Poisson points in the triangle bounded by the horizontal axis, the vertical line at "time" y and the line connecting $(y-s, 0)$ to (y,s). The probability that there are no points in this triangle is $e^{-\lambda s^2/2}$. Hence,

$$EN_s = \int_s^L e^{-\lambda s^2/2} \lambda s dy$$

$$\approx e^{-\lambda s^2/2} \lambda s L.$$

This expression for the expected number of unraveling segments has its maximum at $s^* = \lambda^{-1/2}$ and at this time the expected number of segments is

$$EN_{s^*} = \sqrt{\lambda/e}\ L.$$

Finally, if we relax the assumption that unraveling occurs at rate one half and instead allow it to occur at rate $v/2$ then the only change that must be made is to replace λ with λv and L with L/v. After doing this, we note that all units work out correctly. That is, all the asymptotic expressions for σ_L do indeed have units of seconds. In particular, note that the expression appearing in the logarithms, $L^2\lambda/v$, is a dimensionless quantity.

ACKNOWLEDGEMENT

We thank D. Aldous, B. F. Logan and M. Westcott for helpful conversations bordering on coauthorship. We especially thank M. Westcott for suggesting the problem to us.

REFERENCES

[1] Cramer, H., and Leadbetter, M.R., *Stationary and Related Stochastic Processes; Sample Function Properties and their Applications*, New York: Wiley (1967).

[2] Feller, W., *An Introduction to Probability Theory and Its Applications, Vols. 1 and 2*, New York: Wiley (1966).