AN APPLICATION OF SYMBOLIC COMPUTATION TO A GIBBS MEASURE MODEL

J. Michael Steele†, Princeton University

ABSTRACT

Gibbs models provide a flexible alternative to traditional methods for specifying statistical models. Because of the availability of cheap computation, the Gibbs alternative is now substantially competitive with traditional methods in contexts where one can be satisfied with system simulations, but Gibbs models are at some disadvantage in theoretical contexts because their analytical properties are more difficult to investigate than those of models built from independent variables.

The first purpose of this talk is to present a specific Gibbs model concerning random trees and show how MACSYMA proved of value in examining its theoretical properties. A second purpose is to initiate a broader discussion of the role of MACSYMA in statistical (and other scientific) research.

1. Introduction to Gibbs Models

Most traditional models in statistics are based on functions of independent random variables, or more generally, on variables whose joint distribution is specified by means of a covariance structure. Although such models serve us well in many contexts, there are numerous problems where physical or other phenomena exhibit a qualitative dependency which is not easily expressed by such techniques. It is in such cases that the flexibility provided by the method of Gibbs models can be especially useful.

To illustrate the idea of a Gibbs measure in as concrete a way as possible we will concentrate on measures on finite sets S. In the description of Gibbs models it is useful to employ some evocative language. In particular, the set S will be thought of as the set of possible states of a physical system, and several notions of thermodynamics will guide our notation. The most essential ingredient of the method is the function $f: S \to \mathbb{R}$ which we will call a *potential* function. The function f has several interpretations as either a measure of complexity or of likelihood, but the bottom line is that we use f to define a measure on S by the formula

$$\mu_t(s) = \frac{e^{-f(s)/t}}{Z(t)}. \qquad (1.1)$$

Here, of course, the denominator $Z(t)$ is chosen in just the way needed to make S have total mass one, *i.e.* $\mu_t(S) = 1$ and $Z(t) = \sum_{s \in S} e^{-f(s)/t}$. The choice of t as the indexing parameter reflects the historical origins of μ_t, where t carries the interpretation of temperature. From the perspective of statistics we should note that the change of variable $\theta = 1/t$ brings (1.1) into the form of an exponential family with natural parameter θ. To benefit from the physical intuition behind (1.1), one should note that as $t \to \infty$, the measure μ_t tends to the uniform distribution on S. This limit relation captures the notion that at high temperature any state of S is as likely as any other state. Conversely, as $t \to 0$, the probability measure μ_t becomes increasingly concentrated on the set $\{s: f(s) = \min f(s')\}$, the interpretation of which is that at low temperature a few especially low energy states become extremely likely.

Now, these stories from physics are all well and good but how do they influence probability models which do not have direct physical interpretation? The main point of Steele (1987) is that combinatorial problems can often be associated with an intuitive Gibbs model with a direct combinatorial interpretation. In fact, the Gibbs paradigm does not need much pushing to be seen to be extremely useful in many situations which are quite removed from the usual problems of thermodynamics.

While the preceding remarks seem necessary to motivate our specific problem, the main purpose of the present discussion is to show how MACSYMA was useful in the theoretical exploration of a simple Gibbs model from combinatorics. A second purpose is to review for a statistical audience some of the literature on the use of symbolic computation in applied science.

2. Trees of Prüfer and Rényi

The specific Gibbs model which concerns us has its origin in the hunt for a tractable parametric family of random trees which can serve to model those trees which one finds in problems of computer science. To get some perspective we will recall some results concerning the uniform distribution on random trees.

One of the earliest results in the enumerative theory of combinatorics is the remarkable theorem of Cayley which says there are exactly n^{n-2} labeled trees on n vertices. When Rényi (1959) took up the modeling of random trees, it was natural to consider the uniform model where any one of the Cayley's n^{n-2} labeled trees would have the same probability of being chosen. The object of main interest to Rényi was the number of leaves of a random tree, and to obtain a handle on this problem, Rényi relied on an elegant code for labeled trees due to Prüfer (1918).

Prüfer's simple code is directly applicable to many questions concerning trees. One builds a one-to-one correspondence between a labeled tree on n vertices $P_1, P_2, \ldots, P_n$ and an $(n-2)$-tuple with elements from $\{1, 2, \ldots, n\}$ via the following recipe:

> First we find the leaf P_i with the highest value of the index i. We then delete P_i from the tree and record the index j of the unique vertex P_j to which P_i was adjacent. This process is repeated

on the pruned tree until $n-2$ vertices have been deleted (and two vertices are left remaining).

Such sequence of recorded adjacency indices is now called the Prüfer code for the tree, and one can show without great difficulty that the coding provides a bona fide one-to-one correspondence. As a consequence of Prüfer's coding, the formula of Cayley for the number of trees on n vertices becomes obvious; our code has $n-2$ places and each place can be held by n values, so n^{n-2} is the cardinality of the set of labeled trees.

Although it cuts a good story short, it suffices to say that Rényi used the Prüfer code to show that $T(n,k)$, the number of labeled trees with k leaves and n vertices, must satisfy the polynomial identity

$$\sum_{k=2}^{n-1} \frac{T(n,k)\begin{bmatrix} x \\ n-k \end{bmatrix}}{\begin{bmatrix} n \\ n-k \end{bmatrix}} = x^{n-2}. \qquad (2.1)$$

This elegant expression permits one to make many deductions concerning the number of leaves of a tree chosen at random from the set of all n^{n-2} labeled trees. In particular, if we let L_n denote the number of leaves of a random tree then one can easily obtain an asymptotic formula for the expected value for the number L_n:

$$EL_n \sim n/e. \qquad (2.2)$$

With more effort one can also obtain the variance relation:

$$L_n \sim n(e-2)/e^2. \qquad (2.3)$$

Finally, Rényi obtained the central limit theorem for L_n by means of substituting $x = 1 - it/\sqrt{n}$ into his key identity (2.1). Several technical problems needed to be resolved by Rényi as he moved from (2.1) to the asymptotics of Ee^{itL_n}, but his analytical expertise prevailed.

Now the interesting point comes upon us. Rényi's detailed analysis pertain to a model which is almost certainly too limited to apply to any of the empirical situations of interest! For example, suppose that for your set of observed trees all of the trees have about $n/5$ leaves, instead of n/e leaves. The uniform model would be inapplicable, and it seems to offer no place to turn. The question becomes, how can we obtain a parametric family of trees with a natural parameter closely related to the number of leaves?

The answer is simple using a Gibbs model. We just let S be the set of all labeled tree on n vertices, and we define $f(s)$ to be the number of leaves of $s \in S$. If we set $\theta = t^{-1}$, we can write (1.1) in the form

$$\mu_\theta(s) = e^{-\theta f(s) - \phi(\theta)} \qquad (2.4)$$

where $\phi(\theta) = \log(\sum e^{-\theta f(s)})$ and $\phi'(0) = E_\theta L_n$ is the expected number of a random tree chosen according to the probability measure (2.4).

By varying our choice of θ can make $E_\theta L_n$ range through the whole interval $[2, n-1]$, and thus we have a family of measures on trees which is essentially indexed by the expected number of leaves.

This model has a number of benefits. First, it is an exponential model so many of the nicest results of mathematical statistics apply to the estimation of θ. Also, it is a Gibbs model so the method of Metropolis *et.al.* can be used to generate random trees which satisfy the given law. Finally, as shown in Steele (1987), the random variables L_n are asymptotically normal.

The purpose of the next section is to illustrate how MACSYMA entered into the exploration of this model and how it helped lead to the resolution of the asymptotic normality of L_n under the Gibbs model.

3. Harper's Method and Computational Empiricism

The central limit theorem for L_n was obtained in Steele (1987) through the application of a technique which was introduced by Harper (1967) for the study of Stirling numbers. Harper's basic idea is so simple it can be discussed in a beginning probability course. Still, the method is quite powerful, and, even when it fails, it often brings a new analytical insight into the problem.

So, what is this nifty method? Suppose you wish to study a random variable Y with probability generating function $h(x)$, and you consider the equation:

$$h(x) = p_0 + p_1 x + p_2 x^2 + \ldots + p_n x^n = 0. \qquad (3.1)$$

Now, suppose for some reason you know that (3.1) has only real roots, r_i, $1 \le i \le n$. Since the p_i are non-negative, the r_i are all non-positive, and h can then be factored into the form

$$h(x) = \prod_{i=1}^{n} (x + r_i)/(1 + r_i) = \prod_{i=1}^{n} (p_i x + q_i). \qquad (3.2)$$

Now, with this before us virtually all probability questions about Y are easily resolved, since (3.2) says that Y is equal to the sum of n independent Bernoulli random variables.

Returning to the Gibbs model of random trees, Steele (1987) showed that only a little manipulating is required to establish that (1) Rényi's CLT (2) a Berry-Essen refinement of Rényi's CLT and (3) the CLT for random trees under the general Gibbs model will all follow easily once one proves that the equation

$$\sum_{k=2}^{n-1} T(k,n) x^k = 0 \qquad (3.3)$$

has only real roots.

The handles we have on (3.3) are that $T(n,k)$ can be expressed in terms of Stirling numbers of the second kind by

$$T(n,k) = s(n-2, n-k) n!/k! \qquad (3.4)$$

and the Stirling numbers satisfy the classical recursion

$$s(n,k) = k s(n-1,k) + s(n-1,k-1).$$

Once we learn enough about the geometry of zeros of polynomials, it is not hard to show (3.3) has only real roots; but before even starting to learn what needs to be known, it seems to be almost a psychological requirement to find out whether Harper's method really has a chance. After all, it

is quite possible that (3.3) does have non-trivial complex roots.

Fortunately, one can easily investigate (3.3) using MACSYMA and the following interactive code should illustrate the point. (The c-lines are entered command and the d-lines are MACSYMA's replies. The other notations should either be intuitive, or at least easily understood after reading the companion article by Rand (1987).)

(c1) $s[n,k] := s[n-1,k]*k + s[n-1,k-1];$

(d1) $s_{n,k} := s_{n-1,k}k + s_{n-1,k-1};$

(c2) for n:1 through 50 do ($s[n,1]$:1);

(d2) done

(c3) for n:1 through 50 do ($s[n,n]$:1);

(d3) done

(c5) $t[n,k] := s[n-2,n-k]/\text{factorial}(k);$

(d5) $t_{n,k} := \dfrac{s_{n-2,n-k}}{k!};$

{Technical Remark: One should note that $t_{n,k}$ differ from the values $T(n,k)$ of (3.3) only by a factor of $n!$, a factor not impacting the location of roots.}

(c6) $p[n](x) := \text{sum}(x\hat{}j*t[n,j],j,2,n-1);$

(d6) $p_n(x) := \text{sum}(x^j t_{n,j},j,2,n-1);$

(c7) $p[5](x);$

(d7) $\dfrac{x^4}{24} + \dfrac{x^3}{2} + \dfrac{x^2}{2};$

{Technical Remark: This differs from the probability generating function for L_n by a factor of $n!/n^{n-2}$, which for $n = 5$ is a reassuring 24/25.}

(c8) all roots (%);

(d8) [$x = 0.0$, $x - 0.0$, $x = -1.101020514433644$, $x = -10.89897948556636$];

(c9) $p[8](x);$

(d9) $\dfrac{x^7}{5040} + \dfrac{31x^6}{720} + \dfrac{3x^5}{4} + \dfrac{65x^3}{24} + \dfrac{5x^3}{2} + \dfrac{x^2}{2}$

All roots again can do its job and find that the roots are, to more places than we care to record, {0, 0, −0.27, −0.99, −3.23, −14.20, −198.28}. In the same way, we find for $n = 15$ the roots of $p[15](x)$ are {0, 0, −0.065, , . . . , −56488.80}, and again we find that all of the roots are indeed real.

Before leaving this example, or embarking upon any summary observations about symbolic computation, some points should be made concerning zero locations and their role in Gibbs models. First, we should note that proof of the fact that (3.3) has only real roots is given in Steele (1987), where a key lemma is provided by a result due to Pólya and Schur (1914) which exhibits a polynomial to polynomial transformation which preserves the property of having only real roots. The interplay between Gibbs models and zero location are amusing in the combinatorial context but there are also far more serious contributions with a similar flavor. Kac (1974) tells a charming story how a result on zeros due to Pólya (1926) was instrumental in the work of C.N. Yang and T.D. Lee in the theory of phase transition of a lattice gas.

4. MACSYMA in Applications

The use of MACSYMA reported here has several properties which seem typical of cases where MACSYMA proves to be of real service in mathematical research. First, the assistance provided by MACSYMA came in the form of encouragement rather than proof. Second, MACSYMA became engaged with the problem through the exploration of a significant *example*, rather than by frontal assault on the general problem. Third, MACSYMA provided a convenient interactive environment for performing a calculation which *could* have been performed with far humbler tools.

There is no reason to apologize for any of these properties, and in fact, there is some power in embracing these features as the honest path of MACSYMA application. The problem considered here was amenable to a very straight forward application of MACSYMA as a tool of exploratory mathematics. Still, even in modest exploration, MACSYMA sometimes will buckle under the weight of a seemingly reasonable request. Fortunately, calculations in MACSYMA can be guided toward a problem along many different paths and sometimes such guidance makes a big difference. For a study of such a case and its happy resolution, one should read the discussion provided in Campbell and Gardin (1982) concerning some symbolic determinants.

This report has emphasized the easy uses of MACSYMA, but MACSYMA can indeed support extensive development. The possibilities of such development are well brought out by Watanabe (1984) which reports on the successes and failures of a 1400 line MACSYMA ODE solver as applied to the 542 equations solved in Kamke (1959).

In addition to personal exploration, the best place to learn the capabilities of MACSYMA is Rand's *Computer Algebra in Applied Mathematics: An Introduction to MACSYMA*. The examples in that volume guide one through a useful exploration of MACSYMA's facilities, but its principle benefit is that the focus is kept on the *use* of MACSYMA in scientific work, in contrast to professional involvement with symbolic computation. There are numerous surveys of the application of MACSYMA in science and engineering, and one can get a good feeling for the scope of MACSYMA applications from Engeler and Mäder (1985), Fitch (1979), and Pavelle (1985).

Finally, some comment should be made which acknowledges that MACSYMA is but one of many symbolic computational environments. Loos (1982) mentions that there are at least 60 computer algebra systems. Besides MACSYMA, the best known are probably muMATH, REDUCE, and SCRATCHPAD, and each of these has its

own virtues. One possibly historic development in symbolic computation took place this month with the introduction of the Hewlett-Packard 28C, the first hand-held calculator to do symbolic algebra and symbolic calculus. If history serves as a reliable guide, symbolic algebra will soon be as available as there are people with a desire to participate in its benefits.

† Research supported in part by National Science Foundation Grant #DMS-8414069.

References

Campbell, J.A. and Gardin, F. (1982). "Transformation of an intractable problem into a tractable problem: evaluation of a determinant in general variables," *Computer Algebra: EUROCAM '82* (G. Goos and J. Hartmanis, eds.), Lecture Notes in Computer Science 144, Springer-Verlag, New York.

Engeler, E. and Mäder, R. (1985). "Scientific computation: the integration of symbolic, numeric and graphic computation," in *EUROCAL '85* (G. Goos and J. Hartmanis, eds.), Lecture Notes in Computer Science, 203, Springer-Verlag, New York.

Fitch, J.P. (1979). "The application of symbolic algebra to physics -- a case of creeping flow," in *European Symposium on Symbolics and Algebraic Computation* (G. Goos and J. Hartmanis, eds.), Lecture Notes in Computer Science 72, Springer-Verlag, New York.

Harper, L.H. (1967). "Stirling behavior is asymptotically normal, *Ann. Math. Statis.*, **38,** 410-414.

Kac, M. (1974). "Comments on 'Bemerkung über die Integraldarstellung der Riemannschen ξ-function,'" pp. 424-426 (also *George Pólya: Collected Papers Vol II Location of Zeros*, pp. 243-255, R.P. Boas, ed., MIT Press, Cambridge Massachusetts).

Kamke, E. (1959). *Differential Gleichungen-Lösungsmethoden and Lösungen*, Chelsea, New York.

Loos, R. (1982). Introduction to *Computer Algebra Symbolic and Algebraic Computation*, (eds. Buchberger, B., Collins, G.E., and Loos, R.), Computing Supplementum 4, Springer-Verlag.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). "Equations of state calculation by fast computing machines, *J. Chem. Physics*, **21,** 1087-1092.

Pavelle, R. (1985). "MACSYMA: capabilities and applications to problems in engineering and the sciences," in *EUROCAL '85*, G. Goos and J. Hartmanis, eds., Lecture Notes in Computer Science 203, Springer-Verlag, New York.

Pólya, G. and Szegö, G. (1976). *Problems and Theorems in Analysis II*, Springer-Verlag, New York, p. 45.

Pólya, G. (1926). "Bemerkung über die Integraldarstellung der Riemannschen ξ-function," *Acta Math.*, **48,** 305-317 (also *George Pólya: Collected Papers Vol II Location of Zeros*, pp. 243-255, R.P. Boas, ed., MIT Press, Cambridge Massachusetts).

Pólya, G. and Schur, I. (1914). "Uber zwei Arten von Faktorenfolgen in der Theorie der algebraischen Gleichungen," *J. Reine Angew. Math*, **144,** 89-113 (also *George Pólya: Collected Papers Vol II Location of Zeros*, 100-124, R.P. Boas, ed., MIT Press, Cambridge Massachusetts).

Prüfer, A. (1918). "Neuer Beweis eines Satzes uber Permutationen," *Archiv fur Mathematik und Physik*, **27,** 142-144.

Rand, R.H. (1987). "Computer algebra applications using MACSYMA," in this volume.

Rand, R.H. (1984). *Computer Algebra in Applied Mathematics: An Introduction to MACSYMA*, Pitman Advanced Publishing Program, Boston.

Rényi, A. (1959). "Some remarks on the theory of trees," *MTA Mat. Kut. Int. Kozl.*, **4,** 73-85.

Steele, J.M. (1987). "Gibbs measures on combinatorial objects and the central limit theorem for an exponential family of random trees, *Probability in the Engineering and Information Sciences*, **1,** 47-59.

Stoutemayer, D.R. (1985). "A preview of the next IBM-PC version of muMATH," in *EUROCAL '85*, G. Goos and J. Hartmanis, eds., Lecture Notes in Computer Science 203, Springer-Verlag, New York.

Watanabe, S. (1984). "An experiment toward a general quadrature for second order linear ordinary differential equations by symbolic computation," *EUROSAM '84*, G. Goos and J. Hartmanis, eds., Lecture Notes in Computer Science 174, Springer-Verlag, New York.