

# The Volume of Bitnets

Carlos C. Rodríguez

*The University at Albany, SUNY*  
*Department of Mathematics and Statistics*  
<http://omega.albany.edu:8008/bitnets>

**Abstract.** A bitnet is a dag of binary nodes representing a manifold of probability distributions for a sequence of binary variables. Bitnets are riemannian manifolds of finite volume when Fisher information is used as the metric. I compute the exact volumes for several interesting topologies including the complete graph, the directed line of  $n$  nodes, the exploding star (naive bayes) and its reverse, the collapsing star. I show a fast algorithm for approximating the volumes of general bitnets. Computer experiments show that the average Ricci scalar of bitnets is always half an integer. Typically bitnets have singular boundaries obtained when some of the conditional probabilities for the binary variables are either zero or one. At the singular points the Ricci scalar becomes negative infinity.

## INTRODUCTION

A bitnet, is a regular statistical model for a sequence of binary variables. The joint probabilities of the sequence are efficiently described by a directed acyclic graph (dag) whose vertices are the variables and whose directed edges indicate stochastic influence. Figure (1) shows four examples of bitnets of 3 variables.

The assumption of a network of stochastic influences for the variables (i.e. a bitnet) allows the following useful factorization of joint probabilities,

$$p(x_V) = \prod_{i \in V} p(x_i | x_{pa(i)}). \quad (1)$$

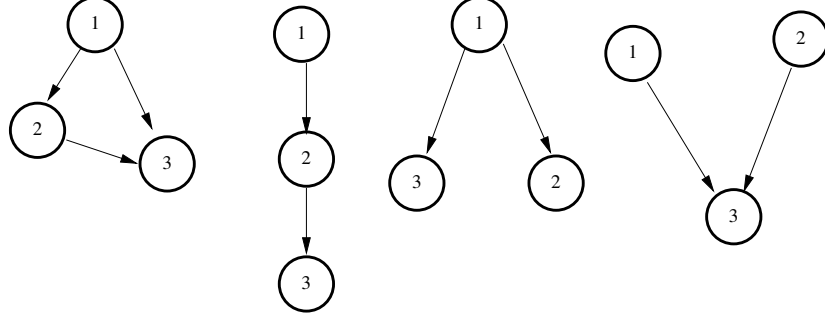
Where  $V$  is the set of vertices of the bitnet,  $pa(i) \subset V$  are the indices of the parents of  $x_i$  and if  $A \subset V$  we denote by  $x_A$  the set of bits with indeces in  $A$ .

It follows from (1) that all the joint probabilities are obtained from the values of  $p(x_i | x_{pa(i)})$ . There are  $2^{|pa(i)|}$  possible values for  $x_{pa(i)}$  and therefore we need that many independent parameters in  $(0, 1)$  for the  $i$ th node. Thus, the total number of independent parameters in  $(0, 1)$  necessary to specify all the joint probabilities of a bitnet is,

$$d = \sum_{i \in V} 2^{|pa(i)|}. \quad (2)$$

For example, the dimensions of the four bitnets in figure (1) are (from left to right): 7, 5, 5, and 6. This paper studies the geometry and topology of the hypothesis space of all the probability distributions of a sequence of binary variables that satisfy the network of stochastic dependencies encoded by a bitnet. By (1), there is a one to one map between

the objects in this set and the points in the  $d$  dimensional open unit cube. We call this kind of hypothesis space a bitnet model.



**FIGURE 1.** Four examples of bitnets. From left to right are: a)  $\bar{K}_3$  the complete dag of 3. b)  $\bar{L}_3$  the directed line of 3. c)  $\bar{E}_3$  the exploding star of 3. d)  $\bar{C}_3$ , the collapsing star of 3.

## THE METRIC OF A BITNET MODEL

When  $|V| = n$  the information metric, (i.e., Fisher information matrix) has  $n^2$  components,

$$g_{ij}(\theta) = E(l_i(X)l_j(X)|\theta) \quad (3)$$

with,

$$l_k(x) = \frac{\partial \log p_\theta(x)}{\partial \theta_k} \quad (4)$$

where  $\theta = (\theta_1, \dots, \theta_d)$  is the vector of parameters in the  $d$ -cube. In order to obtain a simple expression for the metric and the volume of a bitnet, we need to introduce some notation. If  $pa(i) = \{j_1, j_2, \dots, j_k\}$  with  $j_1 < j_2 < \dots < j_k$  then we denote by  $x_{pa(i)}$  both, the set of variables that are parents of  $i$  and the integer obtained when writing the bits that are parents of  $i$  in ascending order, i.e.,

$$x_{pa(i)} = x_{j_1}x_{j_2}\dots x_{j_k} = x_{j_1}2^{k-1} + x_{j_2}2^{k-2} + \dots + x_{j_k}2^0. \quad (5)$$

Also let,

$$\begin{aligned} m(i) &= 2^{|pa(i)|} \\ t(i) &= m(1) + m(2) + \dots + m(i-1) + 1 \\ k(i, x) &= t(i) + x_{pa(i)} \\ v(j) &= \max\{i : t(i) \leq j\} \end{aligned} \quad (6)$$

In words,  $m(i)$  is the number of parameters associated to the  $i$ -th bit,  $t(i)$  is the index of the first parameter of  $i$ . Hence,  $(\theta_{t(i)}, \dots, \theta_{t(i)+m(i)-1})$  is the vector of parameters associated to  $i$ . The integer  $k = k(i, x)$  is such that,

$$p_\theta(x_i|x_{pa(i)}) = \theta_k^{x_i}(1 - \theta_k)^{1-x_i} \quad (7)$$

For example, for  $\vec{K}_3$  (see figure (1)), the parameter  $\theta_5$  represents the probability that bit 3 is on given that bit 1 is off and bit 2 is on. In this case,  $x_{pa(3)} = 01$  which is 1 in decimal. Also  $t(3) = 4$ ,  $k = 4 + 1 = 5$  and  $v(5) = 3$ . Notice that  $v$  is a kind of inverse of  $t$ ;  $v(j)$  gives the bit number  $i$  associated to the parameter  $\theta_j$ .

We can now compute the derivative (4). Taking the logarithm of (1) and using (7) we obtain,

$$l_k = \frac{\partial \log p_{\theta}(x)}{\partial \theta_k} = \begin{cases} \frac{-1}{1-\theta_k} & \text{if } x_i = 0, \text{ when } i = v(k) \text{ and } x_{pa(i)} = k - t(i) \\ \frac{1}{\theta_k} & \text{if } x_i = 1, \text{ when } i = v(k) \text{ and } x_{pa(i)} = k - t(i) \end{cases} \quad (8)$$

Notice that  $i$  is just  $v(k)$ . It appears in the previous formula only to simplify the notation. With (8) we compute  $g_{jk}(\theta)$  by noticing that the product  $l_j * l_k$  has only four possible values,  $\frac{1}{(1-\theta_j)(1-\theta_k)}$ ,  $\frac{-1}{(1-\theta_j)\theta_k}$ ,  $\frac{-1}{\theta_j(1-\theta_k)}$ , and  $\frac{1}{\theta_j\theta_k}$ . Thus, the expected value is just the sum of each value times the probability of obtaining it. It is now straight forward to see that when  $j \neq k$ ,

$$g_{jk}(\theta) = (1 - 1 - 1 + 1)P_{\theta}(x_{pa(v(j))} = a, x_{pa(v(k))} = b) = 0. \quad (9)$$

Where we have used the fact that the probabilities for each of the four cases factorize as a product of two terms when  $j \neq k$  due to the conditional independence assumptions implied by the bitnet model. But when  $j = k$  there is no such factorization and,

$$\begin{aligned} g_{jj}(\theta) &= \left[ \frac{(1-\theta_j)}{(1-\theta_j)^2} - 0 - 0 + \frac{\theta_j}{\theta_j^2} \right] \pi_j(\theta) \\ &= \frac{\pi_j(\theta)}{\theta_j(1-\theta_j)} \end{aligned} \quad (10)$$

where,

$$\pi_j(\theta) = P_{\theta}[x_{pa(v(j))} = j - t(v(j))] \quad (11)$$

Notice that the  $\pi_j$  form  $n$  sequences of probability distributions since, for each  $i = 1, \dots, n$  we have:

$$\sum_{v(j)=i} \pi_j(\theta) = \sum_{j=t(i)}^{t(i)+m(i)-1} \pi_j(\theta) = 1. \quad (12)$$

Since the metric tensor is diagonal, its determinant is the product of the diagonal entries. The total intrinsic volume occupied by a given bitnet is then the integral over the  $d$ -cube of the element of volume. We obtain the general expression,

$$\text{vol}(\text{bitnet}) = \int_{[0,1]^d} \sqrt{\prod_{j=1}^d \frac{\pi_j(\theta)}{\theta_j(1-\theta_j)}} d\theta. \quad (13)$$

Collecting  $j$ 's according to  $v(j) = i$  we can also write,

$$\text{vol}(\text{bitnet}) = \int_{[0,1]^d} \prod_{i=1}^n W_i^{1/2} d\theta \quad (14)$$

where,

$$W_i = \prod_{v(j)=i} \frac{\pi_j(\theta)}{\theta_j(1-\theta_j)} = \frac{\prod_r P[x_{pa(i)} = r]}{\prod_r \theta_{t(i)+r}(1-\theta_{t(i)+r})} = \frac{\prod_r p_i(r)}{\prod_r \rho_i(r)} \quad (15)$$

the index  $r = 0 \dots m(i) - 1$  runs over all the possible values of  $x_{pa(i)}$  (see (5)) and the last equality serves as a definition for  $p_i(r)$  and  $\rho_i(r)$ .

Using the fact that all the  $\pi_j(\theta) \leq 1$  (they are probabilities), the monotonicity of integrals, Fubini's theorem, and the Dirichlet integral:

$$\int_0^1 \frac{dt}{\sqrt{t(1-t)}} = \pi \quad (16)$$

(i.e.,  $\text{Beta}(1/2, 1/2) = \pi$ ) we have,

$$\text{vol}(\text{bitnet}) \leq \pi^d, \quad (17)$$

with equality, if and only if, the bitnet consists of totally disconnected nodes. In that case,  $d = n$ ,  $\pi_j = 1$ , and the bits are independent variables. Thus, all bitnet models are riemannian manifolds of finite volume in the information metric. The upper bound (17) can be improved (see below).

## VOLUMES OF SPECIAL TOPOLOGIES

We now provide exact formulas for the volumes of the four topologies exemplified in figure (1):  $\vec{K}_n, \vec{L}_n, \vec{E}_n, \vec{C}_n$ .

### Complete Dags: $\vec{K}_n$

The complete bitnet of  $n$  nodes has all the available arrows. It has the highest possible dimension,  $d = 2^0 + 2^1 + \dots + 2^{n-1} = 2^n - 1$  for a dag of  $n$ . The hypothesis space  $\vec{K}_n$  is nothing but the multinomial model of the discrete variable whose  $2^n$  outcomes correspond to each possible binary sequence of  $n$  bits. Instead of using the parameters  $\theta = (\theta_1, \dots, \theta_d)$  of conditional probabilities specified by the bitnet, one can always use  $p = (p_0, \dots, p_d)$  with  $\sum p_i = 1$  of the joint probabilities for each sequence of  $n$  bits. This maps the  $d$ -cube into the  $d$ -simplex. We can now map the  $d$ -simplex to the positive part of the  $d$ -sphere of points  $t = (t_0, \dots, t_d)$  with  $t_i > 0$  and  $\sum t_i^2 = 1$  by simply using the standard  $t_i = \sqrt{p_i}$  transformation. It follows immediately from this considerations that the information volume of the complete bitnet is the same as that of the multinomial

which turns out (by using the standard  $\sqrt{p_i}$  transformation) to be exactly half the volume of the unit  $d$ -sphere  $S^d$ , i.e.,

$$\text{vol}(S^d) = \frac{2\pi^{\frac{d+1}{2}}}{\Gamma(\frac{d+1}{2})}. \quad (18)$$

Thus,

$$\text{vol}(\vec{K}_n) = \frac{\pi^{2^{n-1}}}{(2^{n-1} - 1)!} \quad (19)$$

This sequence becomes exponentially small very quickly,

$$\begin{aligned} \text{vol}(\vec{K}_n) &= \pi, \pi^2, \frac{\pi^4}{6}, \frac{\pi^8}{5040}, \dots \sim \frac{1}{\sqrt{2\pi}} \left(\frac{\pi}{k}\right)^k e^{-k} \\ &= 3.14, 9.86, 16.2, 1.87, 0.0000683, \dots \end{aligned} \quad (20)$$

where the asymptotic expression is for  $k = 2^{(n-1)}$  as  $n \rightarrow \infty$ . The Ricci scalar is, not surprisingly, constant and it can be obtained from the corresponding value for the multinomial model:

$$\text{Ricci}(\vec{K}_n) = \frac{d(d-1)}{4} = \frac{(2^n - 1)(2^{n-1} - 1)}{2}. \quad (21)$$

### The Directed Line: $\vec{L}_n$

The  $n$  nodes are connected linearly, i.e., one is the parent of two who is the parent of three, . . . , who is the parent of  $n$ . The dimension is  $d = 1 + (n-1)2 = 2n - 1$ . The exact volume becomes very difficult to compute for values of  $n \geq 4$  but computer experiments carried with the aid of *vTool*<sup>1</sup> show that,

$$\text{vol}(\vec{L}_n) = 4 \left(\frac{\pi}{2}\right)^{3n-4}. \quad (22)$$

The first two cases  $n \in \{1, 2\}$  are trivial (they are also complete bitnets) so the values of  $\pi$  and  $\pi^2$  are immediately obtained. The case  $n = 3$  is already not easy (maple can't do it alone) but a special change of variables in 3D shows that this volume must be the same as the volume of  $\vec{E}_{2+1}$  which is  $\frac{\pi^5}{8}$  (see below equation (24)). Many other values of  $n$  have been estimated by Monte Carlo and they provide strong evidence to the validity of (22). The Ricci scalar for the case  $n = 3$  is computed with *vTool* as,

$$R = 5 - \frac{1}{2\rho} \quad (23)$$

---

<sup>1</sup> *vTool* is a maple module available at <http://omega.albany.edu:8008/bitnets/>.

where  $\rho$  is the variance of the central node. I believe that to hold in general, i.e., that the scalar curvature always depends on the variance of the central nodes. By looking at the components of the Riemann tensor it is possible to show that the scalar curvature is always independent of the parameters of the leave nodes.

### The Exploding Star: $\vec{E}_{n+1}$

One parent node with  $n$  children. This is what the machine learning community calls Naïve Bayes. The children bits (e.g., the observed symptoms) are assumed to be independent conditionally on the parent (the presence or absence of the disease). This bitnet is by far the most popular bayesian network due to its simplicity and to its surprising accuracy and robustness. As I argue below, their volumes can be used to explain, at least in part, their success in applications.

The volume of  $\vec{E}_{n+1}$  is easily obtained from the general formula (14),

$$\text{vol}(\vec{E}_{n+1}) = \int \left[ \frac{1}{\rho_1} \frac{\rho_1}{\rho_2 \rho_3} \frac{\rho_1}{\rho_4 \rho_5} \dots \frac{\rho_1}{\rho_{2n} \rho_{2n+1}} \right]^{1/2} d\theta \quad (24)$$

separating the variables by Fubini, using (16) and defining,

$$B(r) = \int_0^1 t^{r/2} (1-t)^{r/2} dt = \int \rho_i^{r/2} d\theta = \frac{\Gamma(\frac{r}{2} + 1)^2}{\Gamma(r + 1)} \quad (25)$$

we obtain,

$$\text{vol}(\vec{E}_{n+1}) = \pi^{2n} B(n-1). \quad (26)$$

The sequence of volumes explodes exponentially,

$$\begin{aligned} \text{vol}(\vec{E}_{n+1}) &= \pi^2, \frac{\pi^5}{8}, \frac{\pi^6}{6}, \dots \sim \sqrt{2\pi n} \left( \frac{\pi^2}{2} \right)^n \\ &= 9.87, 38.25, 160.24, 698.69, 3121.6, \dots \end{aligned} \quad (27)$$

The computation of the scalar curvature was obtained with the help of *vTool*. It depends only on the variance  $\rho = \rho_1$  associated to the parent (center) node,

$$\text{Ricci}(\vec{E}_{n+1}) = R(\rho) = \frac{n}{2} \left[ (2n+1) - \frac{n-1}{2\rho} \right] \leq \frac{3}{2}n \quad (28)$$

As  $\rho \rightarrow 0$  the scalar curvature diverges:  $R(\rho) \rightarrow -\infty$ . The boundary  $\partial \vec{E}_{n+1}$ , consisting of all the models with coordinates  $\theta \in [0, 1]^d$  with at least one component  $\theta_j = 0$ , contains the surface  $\rho = 0$  of singularities. This has an easy explanation. The estimation of the parameters of the model obviously becomes more difficult when the variance of the parent node is close to zero. If there is only one in a billion chance of observing a given disease, we need billions of observations to estimate the probabilities with a Naïve Bayes

model. In the limit of zero variance we are at an inferential black hole. No information (about *any* of the parameters of the model) can *ever* be extracted from data. Notice that the scalar curvature does not depend on the parameters of leave nodes. That is always the case for all bitnets.

The average scalar curvature,  $\langle R \rangle$ , is computed by integrating  $R$  given by (28), with respect to the volume element of  $\vec{E}_{n+1}$  and dividing by the total volume (26). We obtain:

$$\langle R \rangle = \frac{n}{2}. \quad (29)$$

### The Collapsing Star: $\vec{C}_{n+1}$

This bitnet corresponds to the one child of a promiscuous mother! i.e.,  $n$  parent nodes with only one child. It has dimension  $d = 2^n + n$  and its volume is computed straight forwardly from (14),

$$\text{vol}(\vec{C}_{n+1}) = \int \left\{ \frac{1}{\rho_1} \frac{1}{\rho_2} \cdots \frac{1}{\rho_n} \frac{[\rho_1 \rho_2 \cdots \rho_n]^{2^{n-1}}}{\rho_{n+1} \cdots \rho_{n+2^n}} \right\}^{1/2} d\theta \quad (30)$$

integrating out the parameters of the  $2^n$  leave nodes, using Fubini and the function  $B$  defined in (25) we can write,

$$\text{vol}(\vec{C}_{n+1}) = \pi^{2^n} B(2^{n-1} - 1)^n \quad (31)$$

With the aid of *vTool* we can show that the Ricci scalar has the form,

$$R = a - b \left( \frac{1}{\rho_1} + \frac{1}{\rho_2} + \cdots + \frac{1}{\rho_n} \right) \quad (32)$$

where  $(a, b)$  depend only on  $n$ . The only known values are for  $n = 1, 2, 3, 4$ . They are  $(a, b) = (3/2, 0), (10, 1/2), (54, 3), (272, 14)$  with average Ricci scalars  $\langle R \rangle = 3/2, 2, 6, ?$ . Equation (32) tells us that, as geometric objects, a child node with only one parent is radically different than a child node with two or more parents. When  $n = 1$  the space has constant scalar curvature but when  $n \geq 2$  the curvature diverges to minus infinity as we let the variance of any of the parent nodes to go to zero. So what's so different about the cases  $n = 1$  and  $n = 2$ ? What does curvature really mean statistically? I think what makes the cases  $n = 1$  and  $n = 2$  different is that with only two nodes we can reverse the arrow, obtaining a bitnet which is markov equivalent to the original one (same V structures) but when  $n = 2$  reversing arrows produces a non markov equivalent bitnet. Thus, with only one parent and one child, if the parent has a variance very close to zero the apparent singularity disappears by the reparametrization implied by the reversing of the arrow making the parent a leave node. Being a true geometric invariant, the information contained in the Ricci scalar holds in all possible descriptions (reparametrizations, markov equivalence transformations, etc..) of the model and it must be telling us something significant about the difficulty of estimation at each point.

## BOUNDS FOR GENERAL VOLUMES

For most large bitnets the exact computation of their volumes becomes impractical. This section shows cook-book formulas for a lower and an upper bound for the volumes of general bitnets. These bounds can be shown to be exact for complete bitnets and for bitnets with maximum depth of 1 (i.e., bitnets without grand parents). The geometric mean between the lower and the upper bound has been observed to perform remarkably well in large simulation studies [1].

I claim that, the exact volume  $Z$  of a general bitnet is always bounded between  $L$  and  $U$ , i.e.,  $L \leq Z \leq U$  where, the upper and lower bounds are given in terms of the function  $B$  defined in (25), by

$$L = \prod_{i=1}^n B^{m(i)}(a_i) \quad (33)$$

with

$$a_i = -1 + \sum_{j \in ch(i)} 2^{|pa(j)| - |pa(i)| - 1} \quad (34)$$

and

$$U = \prod_{i=1}^n B^{m(i)}(b_i) \quad (35)$$

with  $b_i$  being the same as (34) except that now the sum does not run over all the children of  $i$ , but only over those children  $j$  of  $i$  for which  $pa(j) \subset pa(i)$ .

## THE IMPORTANCE OF VOLUMES

Why should we care about the volumes of bitnets? One answer is model selection. On the one hand we want our models to be sufficiently large to be able to approximate the true distribution closely. On the other hand we only have limited resources and we need small models that can be handled efficiently. There are many ways to measure the size of a model but, not surprisingly, the information volume is explicitly showing up in the formulas. Consider, for example, the latest expression for the minimum description length (MDL) criterion for model selection [2, 3],

$$\text{MDL} = - \sum_{i=1}^N \log p(y_i | \hat{\theta}) + \frac{d}{2} \log \frac{N}{2\pi} + \log V \quad (36)$$

where  $N$  is the sample size,  $(y_1, \dots, y_N)$  is the observed data,  $\hat{\theta}$  is the MLE of the vector of parameters,  $d$  is the dimension of the model (number of free parameters) and  $V$  is the information volume of the model. The MDL approximates the length of the best possible code for the observed data that can be built with the help of the model. According to the MDL criterion, the best model is the one that allows the shortest possible encoding of the



observations [2]. It so happens that (36) is also the  $o(1)$  approximation to  $-\log P(M|y)$  (see [3, 4]) i.e., minus the logarithm of the posterior probability of the model  $M$ . Thus, on the basis of the observed data alone, with no other prior information at hand, given the choice between two models of the same dimensionality providing the same fit to the data we must prefer the one with smaller volume  $V$ . One must notice however, that the volume only appears as the third term (of order  $1 = N^0$ ) of the asymptotic expansion (as  $N \rightarrow \infty$ ). The first term ( $-\log$  likelihood fit) scales linearly with  $N$ , the second term scales linearly in the number of parameters  $d$  but logarithmically on the sample size  $N$ . Thus, for sufficiently large  $N$ , the volume term will eventually become negligible but exactly *when* it does become negligible depends on the specific models being considered. Simulation studies [1] show improvements in model selection of the order of 28% on the average, when the expression (36) is used instead of the traditional MDL without the volume term.

The modern expression for the MDL (36) exemplifies the natural desirability of models with small volume. Desirability for large volumes is naturally found on measures of generalization power. As just one example consider the expression for the generalization power of heat kernels [5],

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_R(\mathbf{x})) = O\left(\left(\frac{V}{t^{\frac{d}{2}}}\right) \log^{\frac{d+2}{2}}\left(\frac{1}{\varepsilon}\right)\right) \quad (37)$$

where  $\mathcal{N}(\varepsilon, \mathcal{F}_R(\mathbf{x}))$  denotes the size of the smallest  $\varepsilon$ -cover of the space  $\mathcal{F}_R(\mathbf{x})$  which is the ball of radius  $R$  (in the sup-norm) of functions defined on the data  $\mathbf{x}$  in terms of a heat kernel  $K_t$ . The specific technicalities of this method of estimation are not very important for us here. The main point is that for a given accuracy of estimation ( $\varepsilon$ ), between two competing models of the same dimension  $d$ , we must choose the one with the largest information volume  $V$  in order to increase generalization power.

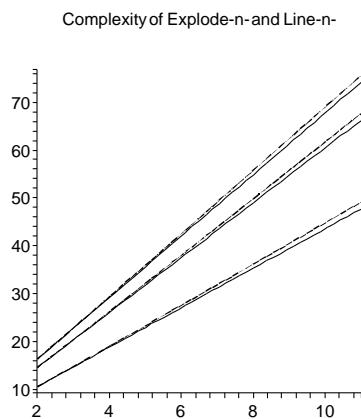
## COMPLEXITY

Let us separate the expression (36) as,

$$\text{MDL} = \text{Fit} + \text{Complexity} \quad (38)$$

where by ‘‘Complexity’’ we simply mean the sum of the last two terms in (36). These are the terms that do not involve the observed data, only their number  $N$ , the dimension of the model  $d$ , and its volume  $V$ . Figure (2) shows the complexity terms for the  $\vec{E}_n$  and  $\vec{L}_n$  bitnets for binary sequences of different lengths  $n < 12$  and for three sample sizes  $N = 100, 500, 1000$ . The picture is clear. The complexities (actually just their volumes for they have the same  $d$ ) of the exploding star and the line are very similar straight lines with the exploding star always above the line. Complexity increases with both, the size of the network  $n$  and the sample size  $N$ . This means that by adding more leaves (symptoms) to a Naïve Bayes network we increase its complexity and by increasing its volume we (probably) increase its power of generalization.

On the other hand figure (3) shows a very different behavior for the  $\vec{C}_n$  bitnet. The complexity reaches a maximum saturation point and then decreases without bound.



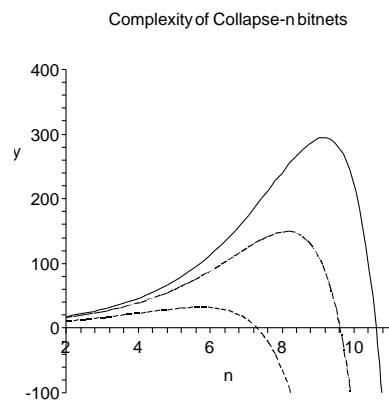
**FIGURE 2.** Complexity of  $\vec{E}_n$  and  $\vec{L}_n$  for different sample sizes  $N = 100, 500, 1000$ . The magnitude of curves increases with  $N$  and Explode  $>$  Line

Thus, adding more parents may help increase generalization power but after crossing the saturation size the bitnet will loose complexity and probably its power to generalize correctly as well.

I do believe that there is more to the story of model complexity than what's available from (36). Recall that (36) is only the first three terms of an asymptotic expansion. The neglected terms contain the model curvatures (see [4]) and by neglecting them we are failing to account for the difficulty of extracting information out of the sample due to the presence of high model curvatures. One may define model complexity in pure geometric terms as  $\langle R \rangle$  and then try to characterize the models of extreme complexity with data and in the vacuum, without data. This type of variational problem has been very successful at describing observational data in physics and a lot is already known about it.

## REFERENCES

1. Lauria, E., *Learning the structure and parameters of bayesian belief networks: An application and a methodology in IT implementation*, Ph.D. thesis, The University at Albany, School of Information



**FIGURE 3.** Complexity of  $\vec{C}_n$  for different sample sizes  $N = 100, 500, 1000$ . The magnitude of curves increases with  $N$

- Science and Policy (2003).
2. Rissanen, J., *IEEE Trans. Info. Thr.*, **42**, 40–47 (1996).
  3. Balasubramanian, V., A geometric formulation of occam’s razor for inference of parametric distributions, Tech. rep., Princeton, Physics PUPT-1588 (1996).
  4. Rodríguez, C., A geometric theory of ignorance, Tech. rep., <http://omega.albany.edu:8008/ignorance> (2002).
  5. Lafferty, J., and Lebanon, G., Diffusion kernels on statistical manifolds, Tech. rep., CMU, School of Computer Science (2004).