REDUNDANT UNITS IN HIGH DIMENSIONS

BY

ART B. OWEN

TECHNICAL REPORT NO. 432

JULY 1993

PREPARED UNDER THE AUSPICES

OF

NATIONAL SCIENCE FOUNDATION GRANT DMS92–04864

DEPARTMENT OF STATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
Redundant Units in High Dimensions

By

Art B. Owen

Technical Report No. 432

July 1993

Prepared Under the Auspices

Of

National Science Foundation Grant DMS92-04864

Department of Statistics
Stanford University
Stanford, California
Redundant Units in High Dimensions
Art B. Owen

Department of Statistics, Sequoia Hall, Stanford University
Stanford CA, 94305

Standard asymptotic statistical theory does not apply to connectionist models with redundant units. This paper explores the predictions of Aldous's Poisson clumping heuristic, for redundant units with a large number of inputs.

A simplified version of the findings is as follows: the probability that a single hidden unit with \(d\) input units explains a large amount of variability in pure noise is \(\delta^{(d-1)/2}\) times the probability for a linear hidden unit. Different methods may be characterized by their value of \(\delta\), which is typically a constant greater than or equal to 1. For some methods \(\delta\) itself grows like \(d^2\). Formulas for \(\delta\) are given for several types of units (sigmoidal, radial basis functions, ramps, wavelets) as a function of a natural distance of the unit from the center of the training data.

Within each of these families of methods are versions that are very susceptible to noise, and versions that are much less susceptible. The noise resistant versions tend to be long tailed, in a sense described in this article.

1 Introduction.

This paper considers the effects of redundant units on feedforward connectionist models with a single hidden layer and a large number of input units.

Section 2 describes the redundant unit problem. Section 3 discusses how the Poisson clumping heuristic of Aldous (1989) may be applied. Section 4 considers what happens when the inputs are high dimensional. A spherical Gaussian distribution for the data in the input units is used since it makes it easy to obtain expressions involving the dimension of the input. Section 5 works out examples for specific methods. Section 6 gives conclusions. Section 7 presents some mathematical and computational details underlying this work.

This paper studies what happens when connectionist models are used on a response that is pure noise. Since real responses are almost never pure noise, we list here some reasons for considering this case.

After a large number of units have been fit, what is left may be approximately pure noise.

A method that is extremely prone to finding structure in pure noise will have to be trained much more carefully than one that is relatively immune.

One is naturally suspicious of apparent structure that is smaller than that which might be found in pure noise. Methods for which noise artifacts are small allow one to report small effects with greater confidence.

Cross validate methods are useful to avoid being fooled by artifacts of the noise. But where noise artifacts are large and numerous, there is a danger that they will function as decoys: the net may converge to noise artifacts instead of real structure, and while the artifacts could get uncovered later, the real structure might never be discovered.

2 Redundant Units.

To fix ideas, suppose that a network is used to approximate the conditional expectation \(\mu(x) = E(Y|X = x)\), where \(Y\) is a continuously varying response variable and \(X\) is a vector of predictor variables. The available data are a training set of \((X_i, Y_i)\) pairs for \(i = 1, \ldots, n\) in which \(Y_i = E(Y|X = X_i) + \epsilon_i\), where \(\epsilon_i\) are independent errors with mean zero and variance \(\sigma^2\).
We suppose that the output unit is linear, and so the model may be written

\[ \mu(X) = \sum_{j=1}^{J} \omega_j \phi_j(X, \theta_j) \]  

(2.1)

where the \( \omega_j \) are real valued parameters and the \( \theta_j \) are vector parameters, all to be learned (estimated) from the training data. By choosing \( \phi_j(X, \theta_j) \) appropriately, one can implement Gaussian radial basis functions, sigmoidal units and other such models.

If (2.1) holds and all \( \omega_j \neq 0 \), then standard statistical asymptotic results may be applied. Under reasonable assumptions, these describe the tendency of the estimated \( \theta_j \) and \( \omega_j \) to approach the true values and the tendency of the in-sample squared error to be optimistically small. Suppose that in truth, \( \omega_j = 0 \) in (2.1). Then the \( J \)'th unit is redundant. The standard statistical asymptotics do not apply to this case. White (1989) discusses the standard asymptotics, mentioning that redundant units behave differently.

It is unlikely in practice that (2.1) should hold exactly for any \( J \). The standard asymptotics describe how \( \omega_j \) and \( \theta_j \) approach the best values for use in (2.1), as \( n \) increases. When a large number \( J \) of units is used, it is reasonable to expect that some of them are approximately, though not exactly, redundant. In this case, redundant unit asymptotics may be informative.

A very simple special case of (2.1), for one dimensional \( X \), is broken line regression (Hinkley, 1969),

\[ \mu(X) = \omega_1 + \omega_2 X + \omega_3 (X - \theta_3)_+ \]  

(2.2)

Here \( Z_+ = \max(Z, 0) \) is the positive part of \( Z \). When the true \( \omega_3 = 0 \),

\[ \mu(X) = \omega_1 + \omega_2 X \]  

(2.3)

and the knot location \( \theta_3 \) is not identifiable. If one estimates model (2.2), unaware that \( \omega_3 = 0 \), the standard asymptotics do not apply. The minimum sum of squared errors for model (2.2) is smaller than that for (2.3), and the difference does not have a \( \sigma^2 \chi^2_{(2)} \) distribution as it would for large \( n \) in standard settings. Hinkley (1969) reports from simulations that about 3 degrees of freedom are appropriate, and Owen (1991) finds some theoretical support for this.

When (2.3) holds and (2.2) is estimated, there is a tendency for the estimated (and spurious) \( \theta_3 \) to be close to the extreme \( X_i \)'s.

3 Poisson Clumping.

Consider the two models

\[ \mu(X) = A(X)\beta \]

(3.1)

\[ \mu(X) = A(X)\beta + \omega \phi(X, \theta) \]

(3.2)

where \( A(X) \) is a row vector of basis functions, \( \beta \) is a column vector of linear coefficients, \( \omega \) is a scalar parameter, \( \theta \) is the vector of \( p \) nonlinear parameters, and \( \phi(X, \theta) \) is a nonlinear function of \( \theta \) for at least some \( X \). Model (3.2) is less general than model (2.1), but the latter can be approximated by incorporating linear approximations to the first \( J - 1 \) units into \( A(X) \).

Now suppose that (3.1) is true. The best fitting model of the form (3.2) will have smaller squared error than the best fitting model of the form (3.1). Let \( S \) be this reduction in squared error. Let \( S(\theta) \) be the reduction in squared error of (3.2) over (3.1) when \( \theta \) is held fixed. Also let
\[ \hat{\omega}(\theta) \] be the corresponding estimate of \( \omega \). Then \( S = \sup_{\theta \in \Theta} S(\theta) \) where \( \Theta \) is the set of \( \theta \) values searched over. Standard asymptotics apply to \( S(\theta) \) and \( \hat{\omega}(\theta) \) for any fixed \( \theta \), but not to \( S \) itself.

The Poisson clumping heuristic of Aldous (1989) provides an approach to the study of \( S \). Let \( Z(\theta) = \pm S(\theta)^{1/2} \) where the sign of \( Z(\theta) \) matches that of \( \hat{\omega}(\theta) \). Then for large \( n \), or nearly Gaussian \( \epsilon_i \), \( Z(\theta) \) is approximately a smooth Gaussian process, assuming \( \phi(X, \theta) \) is twice differentiable with respect to \( \theta \). The mean of \( Z \) is everywhere 0, the variance is \( \sigma^2 \) and \( \rho(\theta, \vartheta) = \text{cor}(Z(\theta), Z(\vartheta)) \) is the correlation function.

Using Chapters A and J of Aldous (1989), Owen (1993) finds that

\[ P(S > \sigma^2 Z_0^2) \approx 2(2\pi)^{-p/2} Z_0^{-1} \varphi(Z_0) \int_{\theta \in \Theta} |\Lambda(\theta)|^{1/2} d\theta, \quad (3.3) \]

where \( \Lambda \) is the matrix with \( rs \) element

\[ \Lambda_{rs} = -\frac{\partial^2 \rho(\theta, \vartheta)}{\partial \theta_r \partial \theta_s} |_{\theta = \vartheta}, \]

and \( |\Lambda| \) is the determinant of \( \Lambda \). Here \( \varphi(z) = (2\pi)^{-1/2} e^{-z^2/2} \) is the standard univariate Gaussian density. For \( \vartheta \) near \( \theta \),

\[ \rho(\theta, \vartheta) = 1 - (\vartheta - \theta)' \Lambda(\vartheta - \theta)/2. \quad (3.4) \]

Approximation (3.3) is most accurate when \( Z_0 \) is large, \( \Theta \) is large, \( \rho \) shows weak long range dependence and most of \( \Theta \) is not near its boundary \( \partial \Theta \). Formula (3.3) suggests that spurious hidden units are most likely to arise in regions of \( \Theta \) with large values of \( |\Lambda| \). Though (3.3) has many caveats attached, there are intuitive reasons why large \( |\Lambda| \) should indicate regions likely to have spurious units, even when (3.3) does not provide accurate estimates of \( P(S > \sigma^2 Z_0^2) \). Consider two regions \( \Theta_1, \Theta_2 \subset \Theta \), of identical size and shape. If \( |\Lambda| \) tends to be larger in \( \Theta_1 \) then the correlations \( \rho(\theta, \vartheta) \) decrease faster there as \( |\theta - \vartheta| \) increases. Where correlations are small, sharper extremes are to be expected. Another intuitive explanation is that \( \Lambda^2 \) is the covariance matrix of \( \nabla_\theta Z(\theta) \). Where \( |\Lambda| \) is large, \( Z \) tends to be steep and hence hits more extreme levels.

For broken line regression, \( |\Lambda| \) is largest for knots near the edges of the data range. Special one dimensional methods may be applied to approximate \( \sup_{[a,b]} Z(\theta) \) for this problem. Large \( Z(\theta) \) events may be separated into those due to large \( Z(a) \) and those due to "upcrossings" by \( Z \) of high levels somewhere between \( a \) and \( b \). Formula (3.3) captures the upcrossing part, neglecting the boundary part at \( Z(a) \).

For broken plane regression, Owen (1993) finds that \( |\Lambda| \) is largest for bends near the convex hull of the \( X_i \). Similarly, for Gaussian radial basis functions \( |\Lambda| \) is large for Gaussians with small variance parameters and in locations, such as gaps in the \( X_i \), where more than one potential outlier can be explained. Sigmoidal units with large \( |\Lambda| \) tend to have lots of data points on their steep portions.

Owen (1993) gives an expression for \( \Lambda \). Let \( A \) be the matrix with \( i \)'th row \( A(X_i), i = 1, \ldots, n \). Without loss of generality, suppose that \( A \) is of full rank. (Otherwise drop some columns.) Let \( M = A(A'A)^{-1} A' \) be the projection matrix on \( A \), with \( M \) equal to the zero matrix if \( A(X_i) \) has no basis functions. Let \( \phi \) be the vector of \( \phi(X_i, \theta) \) values, \( \phi_r \) the vector \( \partial \phi(X_i, \theta)/\partial \theta_r \), \( < g, h > = g'(I - M)h/n \) and \( \gamma_r = < \phi_r, \phi > / < \phi, \phi > \). Owen (1993) proves the following

**Theorem.** In the above notation

\[ \Lambda_{rs} = < \phi_r, \phi_s > / < \phi, \phi > - < \phi_r, \phi > < \phi_s, \phi > / < \phi, \phi >^2 \]

\[ = < \phi_r - \gamma_r \phi, \phi_s - \gamma_s \phi > / < \phi, \phi >. \quad (3.5) \]
Due to fortuitous cancellation, second derivatives of $\phi$ do not appear in (3.5).

4 Dimension Effect.

In this section we use the heuristic to investigate the effects of changing dimension of $\theta$. For the input training data $X_i \in \mathbb{R}^d$, $d \geq 2$, we consider a spherical Gaussian point cloud with very large $n$. The Gaussian distribution is convenient here since it is the only distribution for which the components of $X$ are independent and the density of $X$ is radially symmetric.

Since $n$ is large, we can approximate $n^{-1} \sum_{i=1}^{n} \phi(X_i, \theta)$ by

$$
E_X(\phi(X, \theta)) = \int_{\mathbb{R}^d} \phi(X, \theta)(2\pi)^{-d/2}e^{-X'X/2}dX
$$

and similarly for functions $g(X), h(X)$, we approximate $<g, h>$ by

$$
E_X(gh) - E_X(gA)(E_X(A'\Lambda))^{-1}E_X(A'h).
$$ (4.1)

We consider three choices for $A(X)$, namely the constant $A_C(X) = (1)$, the linear $A_L(X) = (1, X')$ and the simplest case $A_0$ which has zero components, in which case (3.1) is understood to imply $\mu(X) = 0$. Corresponding to these we have

$$
<g, h>_0 = E_X(gh)
$$

$$
<g, h>_C = E_X(gh) - E_X(g)E_X(h)
$$

$$
<g, h>_L = E_X(gh) - E_X(g)E_X(h) - E_X(X'g)'E_X(X'h).
$$

We explore formula (3.5), for projection units, such as sigmoids, and then for radial basis units, such as Gaussians.

4.1 Projection Units. Let $\Theta = \{ \theta \in \mathbb{R}^d : \theta'\theta = 1 \}$ and

$$
\alpha = \alpha(X, \theta) = X'\theta - \rho.
$$

From here on, $\rho \geq 0$ is a radius giving the distance of the unit from the center of the data, and should not be confused with the correlation function of $Z$. For projection type units $\phi(X, \theta) = \xi(\alpha)$ for some function $\xi$, and $\partial\phi(X, \theta)/\partial\alpha = \xi'(\alpha)X^r$ where $\xi'$ is the derivative of $\xi$ and $X^r$ is the $r$th component of $X$. The type of projection unit (e.g. sigmoidal) is implemented through the choice of $\xi$.

$\Theta$ is $d - 1$ dimensional and has no boundary. By symmetry the $d - 1$ by $d - 1$ matrix $\Lambda(\theta)$ is constant over $\Theta$ for any projection unit and any of the three choices of $A$. Therefore

$$
\int_{\theta \in \Theta} |\Lambda(\theta)|^{1/2}d\theta = \frac{2\pi^{d/2}}{\Gamma(d/2)}|\Lambda(\theta_0)|^{1/2},
$$

the surface measure of the unit sphere in $d$ dimensions times $|\Lambda|^{1/2}$ at any point $\theta_0 \in \Theta$.

For convenience let $\theta_0 = (0, \ldots, 0, 1)'$. We need to find $\Lambda_{rs}$ for $r, s = 1, \ldots, d - 1$. For any projection unit and any of the three basis functions $A$ above, $\gamma = 0$, $r = 1, \ldots, d - 1$. Therefore

$$
\Lambda_{rs} = <\phi, \phi> - <\phi_r, \phi_s>. \quad \text{Also if } r \neq s, \text{ then } <\phi_r, \phi_s> = 0.
$$

By symmetry $<\phi_r, \phi_r>$ is constant in $r$, and so, letting $\delta = <\phi, \phi> - <\phi_r, \phi_r>$ we have

$$
\int_{\theta \in \Theta} |\Lambda(\theta)|^{1/2}d\theta = \frac{2\pi^{d/2}}{\Gamma(d/2)}\delta^{(d-1)/2}.
$$ (4.2)
After some calculations, we find that for $A_0$,

$$\delta = \delta_0 = \frac{\int_{-\infty}^{\infty} \xi'(z - \rho)^2 \varphi(z)dz}{\int_{-\infty}^{\infty} \xi(z - \rho)^2 \varphi(z)dz},$$  

(4.3)

for $A_C$,

$$\delta = \delta_C = \frac{\int_{-\infty}^{\infty} \xi'(z - \rho)^2 \varphi(z)dz}{\int_{-\infty}^{\infty} \xi(z - \rho)^2 \varphi(z)dz - \left(\int_{-\infty}^{\infty} \xi(z - \rho)\varphi(z)dz\right)^2},$$  

(4.4)

and for $A_L$,

$$\delta = \delta_L = \frac{\int_{-\infty}^{\infty} \xi'(z - \rho)^2 \varphi(z)dz - \left(\int_{-\infty}^{\infty} \xi'(z - \rho)\varphi(z)dz\right)^2}{\int_{-\infty}^{\infty} \xi(z - \rho)^2 \varphi(z)dz - \left(\int_{-\infty}^{\infty} \xi(z - \rho)\varphi(z)dz\right)^2 - \left(\int_{-\infty}^{\infty} z\xi(z - \rho)\varphi(z)dz\right)^2}. $$  

(4.5)

As before $\varphi(z) = (2\pi)^{-1/2}e^{-z^2/2}$.

The formulas for $\delta$ are simpler than might have been expected. While $\delta$ is defined through $d$ dimensional integrals, $\delta$ reduces to functions of one dimensional integrals and does not even depend on $d$.

4.2 Radial Basis Units. A similar development holds for radial basis units. Let

$$\alpha = \alpha(X, \theta) = (X - \rho\theta)'(X - \rho\theta).$$

For radial basis units $\phi(X, \theta) = \xi(\alpha)$, and $\partial \phi(X, \theta) / \partial \theta_r = \xi'(\alpha)(-2\rho)(X_r - \rho \theta_r)$. The type of radial basis unit (e.g. Gaussian) is determined by the choice of $\xi$.

As with projection units, $\Lambda(\theta)$ is constant over $\theta \in \Theta$. Also, as for projection units, $\gamma_r = 0$ for $r = 1, \ldots, d - 1$, and $r \neq s$ implies $\langle \phi_r, \phi_s \rangle = 0$. It follows that (4.2) holds with $\delta = \langle \phi, \phi \rangle^{-1} \langle \phi_r, \phi_r \rangle$.

The formulas for $\delta$ are not as simple as those for projection units. Let $U, V, W$ be independent random variables with $U$ and $V$ having the standard normal density and $W$ having the $\chi^2_{d-2}$ distribution. The $\chi^2_{(\nu)}$ density is $e^{-y/2}y^{\nu/2-1}/(2^{\nu/2}\Gamma(\nu/2))$ on $(0, \infty)$.

After some calculations, and letting $\alpha = \alpha(U, V, W) = (U - \rho)^2 + V^2 + W$, we get

$$\delta_0 = 4\rho^2 \frac{E(\xi'^2(\alpha)V^2)}{E(\xi(\alpha)^2)},$$  

(4.6)

$$\delta_C = 4\rho^2 \frac{E(\xi'^2(\alpha)V^2)}{E(\xi(\alpha)^2) - (E(\xi(\alpha))^2)^2},$$  

(4.7)

and

$$\delta_L = 4\rho^2 \frac{E(\xi'^2(\alpha)V^2) - (E(\xi'(\alpha)V^2))^2}{E(\xi(\alpha)^2) - (E(\xi(\alpha))^2)^2 - (E(\xi(\alpha)V))^2}. $$  

(4.8)

Unlike projection units, $\delta$ can depend on $d$, through $W$.

Since $V^2 + W$ is $\chi^2_{d-1}$, some of the expectations above can be reduced from triple to double integrals. Note that the joint density of $U, V, W$ multiplied by $V^2$ is again a probability density so that the numerator of $\delta_L$ is a variance, and hence is nonnegative.
4.3 Discussion. Substituting (4.2) into (3.3) with \( p = d - 1 \), the dimension of \( \Theta \), we find that
\[
P(S > \sigma^2 y) = 2(2\pi)^{-(d-1)/2} y^{(d-2)/2} \varphi(y) \frac{2\pi^{d/2}}{\Gamma(d/2)} \delta^{(d-1)/2}
\]
\[
= \frac{2^2 - d}{2} y^{d/2 - 1} e^{-y^2/2} \delta^{(d-1)/2}.
\] (4.9)

Minus the derivative of (4.9) with respect to \( y \) may be recognized as a linear combination of \( \chi^2 \) densities which integrate back to
\[
P(S > \sigma^2 y) \equiv 2\delta^{d-1} \left( P(\chi^2_{(d)} > y) - P(\chi^2_{(d-2)} > y) \right). \quad (4.10)
\]

For very large \( y \), the \( \chi^2_{(d-2)} \) part of (4.10) can be ignored. Equation (4.10) shows the critical role of \( \delta \). For \( \delta = 1 \), the tail of the distribution of \( S \) is reasonably close to what it would be for a linear unit. For \( \delta > 1 \), a dimension effect takes over in which the connectionist model can find more structure in high dimensional noise than a linear unit can. For \( \delta < 1 \) the connectionist unit finds less structure in high dimensional noise than would a linear unit.

In the next section, we compute \( \delta \) for several types of unit as a function of the radius \( \rho \). The radii of interest must also increase with \( d \), since for large \( d \) most of the \( X \)'s have \( X'X \) close to \( d \). Small \( \rho \), less than 2.5 or so, are of interest for all \( d \), and \( \rho \) near \( d^{1/2} \) are of interest if one allows units to "respond" to a small number of data points.

We evaluate \( \delta(\rho) \) for \( \rho \) between 0 and 35 for some of the examples considered in sections 4.1 and 4.2. For \( d = 1000 \), 99.9\% of the \( X_i \) mass has \( (X'X)^{1/2} < 33.8 \) so the upper limit 35.0 should be ample for \( d \leq 1000 \). For \( d = 100 \), 99.9\% of the \( X_i \) are within 12.2 of the origin and for \( d = 10 \) this radius is 5.4.

5 Examples.

Here we consider various choices for the \( \xi \) function for projection units and for radial basis function units. We find \( \delta \) as a function of \( \rho \) and, in some cases, an auxiliary scale variable \( \tau > 0 \). We choose not to subsume \( \rho \) and \( \tau \) into \( \theta \) because they are likely to suffer significant boundary effects, and because it is interesting to know how \( \rho \) and \( \tau \) affect \( S \). As in section 4 we define \( \alpha = X'\theta - \rho \) for projection units and \( \alpha = (X - \rho \theta)'(X - \rho \theta) \) for radial basis units.

Many useful projection units may be defined via a probability density function \( f(x) \). Let \( F(z) = \int_{-\infty}^{z} f(x)dx \) be the corresponding distribution function, and let \( F(z) = \int_{-\infty}^{z} F(x)dx \) when it exists.

For sigmoidal units \( \xi(\alpha) = F(\alpha/\tau) = F((X'\theta - \rho)/\tau) \). A common choice is the logistic function \( F(z) = e^z/(1 + e^z) \). Units of the form \( \xi(\alpha) = \tau^{-1} f(\alpha/\tau) \) provide activations resembling speed bumps. Units of the form \( \xi(\alpha) = \tau F(\alpha/\tau) \) are ramp functions. For \( \tau > 0 \) they resemble smoothly folded paper and as \( \tau \to 0 \), these folds become creases: \( \xi(\alpha) = \alpha_+ = \max(0, \alpha) \). A convenient fold function is taken from an hyperbola \( \xi(\alpha) = \frac{\alpha}{2} + \left( \frac{\tau^2 + \alpha^2}{4} \right)^{1/2} \). The second derivative of the hyperbolic fold is the density of a \( t_1 \) random variable divided by \( \tau^2 \). Further integrating fold functions, gives functions that approximate \( \alpha_+^p \) for integers \( p \geq 2 \). First or second derivatives of appropriate \( f \) have properties similar to those functions used as wavelets. When a projection unit is described as having light or heavy tails, it means that the underlying density \( f(z) \) has such tails.

For the logistic sigmoid (4.3) becomes
\[
\delta_0 = \tau^{-2} \int e^{1/z - \tau} (1 + e^{1/z - \tau})^{-4} \varphi(x)dx \int e^{1/z - \tau} (1 + e^{1/z - \tau})^{-2} \varphi(x)dx.
\]
So $\delta_0 \leq \tau^{-2}$, and $\delta_0 \rightarrow \tau^{-2}$ as $\rho \rightarrow \infty$. Figure 1 shows $\delta_0$, $\delta_C$, and $\delta_L$ for logistic sigmoidal units with $\tau = 0.5$, for $\rho$ between 0 and 35.0 (inclusive). We note some asymptotes: $\lim_{\rho \rightarrow \infty} \delta_C = \tau^{-2} e^{1/\tau^2}/(e^{1/\tau^2} - 1)$, $\lim_{\rho \rightarrow \infty} \delta_L = \tau^{-2} (e^{1/\tau^2} - 1)/(e^{1/\tau^2} - 1 - 1/\tau^2)$, $\lim_{\tau \rightarrow \infty} \lim_{\rho \rightarrow \infty} \delta_C = 1$, $\lim_{\tau \rightarrow \infty} \lim_{\rho \rightarrow \infty} \delta_L = 2$, $\lim_{\tau \rightarrow 0} \tau^2 \lim_{\rho \rightarrow \infty} \delta_C = 1$, $\lim_{\tau \rightarrow 0} \tau^2 \lim_{\rho \rightarrow \infty} \delta_L = 1$.

The tendency of logistic sigmoidal units to find structure in noisy data is not very sensitive to the radius $\rho$ but is quite sensitive to the steepness $\tau^{-1}$.

For a Gaussian radial basis function, $\xi(\alpha) = e^{-\alpha^2/(2\tau^2)}$ where $\alpha = (X - \rho \theta)'(X - \rho \theta)$ and $\tau > 0$ is a scale parameter. It can be shown that

$$\delta_0 = \frac{\rho^2}{\tau^4 + 2\tau^2}$$ \hspace{1cm} (5.1)

and

$$\delta_C = \delta_0 \left(1 - e^{-\left(\frac{\rho^2}{\tau^2 + 2(\tau^2 + 1)}\right)} \left(1 - \left(\tau^2 + 1\right)^{-\frac{3}{2}}\right)^{-1}\right).$$ \hspace{1cm} (5.2)

So $\delta_0$ turns out not to depend on $d$ and $\delta_C$ depends only weakly on $d$ when $\rho$ is large. The same is true of $\delta_L$, though the expression is more complicated.

For Gaussian radial basis functions, $\delta_0$ increases quadratically with $\rho$ for any $\tau > 0$. So, the Poisson clumping heuristic predicts that Gaussian radial basis functions are much more likely to find structure in pure noise when $\rho$ is large. They are also sensitive to small $\tau$.

Figure 2 shows $\delta_0$, $\delta_C$ and $\delta_L$ for a crease projection, $\xi(\alpha) = \alpha_+$. As $\rho \rightarrow \infty$, all of these $\delta$'s are asymptotic to $\delta_A = (\rho^2 + 5)/2$ which is also plotted. Like Gaussian radial basis functions, crease projection units have $\delta$ increasing as $\rho^2$. For power units, $\xi(\alpha) = \alpha^x$, the asymptote is $\delta_A = \tau(\rho^2 + 4\tau + 1)/(4\tau - 2)$. It is perhaps surprising that the coefficient of $\rho^2$ decreases with $\tau$.

A sigmoidal unit obtained by integrating the uniform $[-\tau, \tau]$ density is piecewise linear. In a personal communication, J. Friedman indicates that such sigmoides have computational advantages over logistic ones. For such units, $\delta$ is asymptotic to $((\rho - \tau)^2 + 5)/2$, which is essentially the same as for crease units. For fixed $\tau > 0$ and large $\rho$ the Gaussian density drops off so quickly as $z$ increases from $\rho - \tau$ to $\rho + \tau$ that the contribution to $\delta$ from $z > \rho + \tau$ is negligible.

The Cauchy distribution, with $f(z) = \pi^{-1}(1 + z^2)^{-1}$ and $F(z) = 1/2 + \tan^{-1}(z)/\pi$, is very long tailed. Figure 3 shows $\delta_C$ for Cauchy sigmoidal units $\xi(\alpha) = F(\alpha/\tau)$ with $\tau$ ranging from 4 to 1/8. Remarkably, for large $\rho$ all these $\delta_C$ curves tend to 1.0. This contrasts with the $\tau^{-2}$ asymptotes for logistic sigmoides. For Cauchy sigmoidal units, $\delta_L$ tends to $2.0$ as $\rho \rightarrow \infty$, and $\delta_0 \propto \rho^{-2}$.

Similarly, for hyperbolic fold functions, $\delta_C \rightarrow 1$, $\delta_L \rightarrow 2$ and $\delta_0 \propto \rho^{-2}$. Indeed if $\xi$ derives from a heavy enough tailed density $f$, one can substitute Taylor expansions of $\xi$ and $\xi'$ around $z = 0$ in the integrands of (4.3), (4.4), (4.5) and find the limits 0, 1, 2 respectively as $\rho \rightarrow \infty$.

We close this section by considering some heavy tailed radial basis functions, that also lead to small values for $\delta$.

For multiquadric radial basis functions

$$\xi(\alpha) = (\alpha + \tau^2)^{1/2}.$$

As $\tau \rightarrow 0$, these become linear (cone shaped) units

$$\xi(\alpha) = \alpha^{1/2}.$$

Figure 4 shows $\delta_C$ for $d = 20, 60, 100$ for multiquadric radial basis functions. The range of $\rho$ in Figure 4 is reduced, since $d \leq 100$ for this plot. The points are obtained numerically, and the curves
are given by the asymptotic formula \( \delta_C \approx (1 + d/2\rho^2)^{-1} \). As \( \tau \) varies from 4 to 0, the numerical values of \( \delta_C \) are visually almost indistinguishable in a plot.

Radial basis functions of the form \( \xi(\alpha) = (\tau^{-2} + \alpha)^{-1} \), which may seem more local than multiquadric radial basis functions, also have \( \delta_C \approx (1 + d/2\rho^2)^{-1} \) for large \( \rho \).

5.1 Illustration. Formula (4.10) says, roughly, that the right tail of the \( S \) distribution is \( \delta^{(d-1)/2} \) times as heavy for a connectionist unit as for a linear one. For \( \delta \geq 1 \), this can be translated into a statement of how much larger \( S \) might be. Suppose that \( d = 30 \). For a logistic sigmoid with \( \tau = 0.5 \), \( \delta \approx 4 \). So the tails of the \( S \) distribution are roughly \( 4^{14.5} \approx 5.4 \times 10^8 \) times as heavy as for \( \chi^2_{(30)} \). This means that there is a 5% chance that \( S > 108.1\sigma^2 \), since \( P(\chi^2_{(30)} > 108.1) \approx 0.05/4^{14.5} \). By comparison the usual asymptotics lead to a 5% chance that \( S > 43.7\sigma^2 \). For a unit with \( \delta = \rho^2/2 \) and \( \rho^2 = d \), the tails are \( 15^{14.5} \approx 1.13 \times 10^{17} \) times as heavy and \( S > 156\sigma^2 \) with probability about 5%.

It is important to note that the probability ratios \( \delta^{(d-1)/2} \) take much more extreme values than do the 95th percentiles of \( S \), although the latter do show substantial differences.

When \( \delta < 1 \), \( S \) has a lighter tail than under standard asymptotics. But it may not be possible to use equation (4.10) to approximate quantiles of \( S \). For example if the tails of \( S \) are .01 times as heavy as those of \( \chi^2_{(d)} \) we cannot find the 95th percentile of \( S \) by solving for \( y \) such that \( P(\chi^2_{(d)} > y) = 0.05/0.01 = 5 \).

6 Conclusions.

It seems reasonable that the best connectionist model to use in an application is one for which the mean response can be approximated accurately with a small number of units. One can approximate one type of unit by several of another type of unit, but whichever type is closest to the function to be learned, has a strong advantage. Thus one might choose sigmoids, ramps or radial basis functions on the basis of which was thought to better fit the structure in the problem at hand. In a family of units thought to be good approximations to the signal, there can be large variations in the degree of noise sensitivity.

Units that approach their asymptotes more slowly, appear to be much less sensitive to noise than are short tailed units of the same general shape. Thus the usual sigmoids are less sensitive to noise than are piece-wise linear sigmoids, and hyperbolic branches are less sensitive than ramps. Similarly, radial basis units that taper off more slowly are less sensitive to noise than very localized ones.

Presumably what matters is how heavy the tails of the units are compared to those of the distribution generating the predictor variables.

Standard statistical asymptotics do not apply to the redundant unit problem. The Poisson clumping heuristic provides some indications, including a simple formula (4.10), for the probability of finding large structures in noise. The heuristic has enough caveats attached to it that these conclusions must be considered tentative until verified by simulation or sharper asymptotics.

Formula (4.10) is most accurate for observations measured with Gaussian errors, and it describes the global minimum of the sum of squared errors. If the error distribution has heavier than Gaussian tails, we would expect (4.10) to underestimate the probability of finding structure. Some extremely noise sensitive methods are ones that are very localized and it may be hard to minimize the corresponding squared errors in practice, so (4.10) may overestimate the noise sensitivity of real algorithms, for short tailed units.

One surprise in this work was that some of the heavy tailed connectionist methods have asymptotics quite similar to the standard ones. This suggests that methods such as Akaike's Information Criterion might be suitable.
7 Technical Details.

Here we summarize some mathematical and computational details omitted from the discussion.

The curves in Figures 1, 2, and 3 were obtained by numerical integration using a Gauss rule with 10 points per "panel" and 200 panels of width 0.5. A panel boundary occurs at \( z = \rho \) so that the lack of smoothness of the integrands for Figure 2 does not result in a great loss of accuracy. The points in Figure 4 were obtained by a product of three Gauss rules, one for each of \( U, V, \) and \( W \) that appear in section 4.2. For \( U, \) 20 panels of width 2 were used. For \( V > 0, \) 10 panels of width 2 were used, the results for \( V < 0 \) being obtainable by symmetry. For \( W, \) 6 panels were used that spanned the interval from 0 to \( q \) where \( P(x_{d-2}^2 > q) \approx 2 \times 10^{-8}. \)

In section 4, when \( d = 2, x_{d-2}^2 \) means a random variable that is always equal to zero.

The fold function \( F \) does not exist for densities that are so heavy that they cannot be twice integrated. For example, there is no Cauchy \( (t_{11}) \) fold function, although the \( t_{22} \) density gives rise to the hyperbolic fold function.

For crease units, \( \xi \) is not sufficiently differentiable. One can find \( \delta \) by approximating the crease by the uniform \([-\tau, \tau]\) density twice integrated, and then letting \( \tau \to 0. \) The Lebesgue dominated convergence theorem (Royden (1968), page 229) applies to the sequence of \( \delta \)'s. A similar approach works for piecewise linear sigmoid.

For sigmoidal unit, \( \alpha^+ \) is obtained by integrating a density once, letting \( \tau \to 0 \) leads to a step function. The limit for \( \delta \) diverges to \( +\infty \) as is appropriate: the resulting \( Z \) process has a covariance that does not satisfy (3.4). (A version of the Poisson clumping heuristic applies to such processes, using the first derivative of the correlation function. See Aldous (1989, Formula J10k).)

The limits as \( \delta \to \infty \) for sigmoidal projection units may be obtained by multiplying numerator and denominator by \( e^{2\alpha^+\tau} \) and then letting \( \rho \to \infty \) in the integrands. These \( \delta(\infty) \) formulas may be Taylor expanded with respect to \( \tau \) to get the limits for large and small \( \tau \).

Formulas (5.1) and (5.2) may be found by completing the square in the exponent of the Gaussian integrals.

For the crease and power units, \( \alpha^+ \), one can apply l'Hôpital's rule to get the limits of \( \delta \), as \( \rho \to \infty \). Multiply the numerator of \( \delta \) by \( \rho^{-2} \) and differentiate numerator and denominator \( 2\delta + 1 \) times.

It should be kept in mind that limits of \( \delta \) as \( \tau \to 0, \rho \to \infty \) depend in general on the order in which they are taken. For example, with fold units obtained from the uniform \([-\tau, \tau]\) density,

\[
\lim_{\rho \to \infty} \lim_{\tau \to 0} \delta_0 / \rho^2 = 1/2 \tag{7.1}
\]

and

\[
\lim_{\tau \to 0} \lim_{\rho \to \infty} \delta_0 / \rho^2 = 1/3. \tag{7.2}
\]

In (7.1), as \( \tau \to 0, \) \( \delta_0 \) for the fold tends to \( \delta_0 \) for a crease. For large \( \rho \) this behaves as \( (\rho^2 + 5)/2. \)

In (7.2), as \( \rho \to \infty \) for any \( \tau > 0, \) \( \delta_0 \) behaves like it would for a unit with \( \xi(\alpha) = \alpha^+ \), and that \( \delta \) is asymptotic to \( (\rho^2 + 5)/3. \)

Acknowledgements.

This work was supported by National Science Foundation grant DMS-9011074.
References.


Legends.

Figure 1.
A logistic sigmoidal projection unit is of the form $e^{\frac{X^\prime - \rho}{\tau}}/(1 + e^{\frac{X^\prime - \rho}{\tau}})$. The three curves are $\delta_0$, $\delta_C$ and $\delta_L$ from formulas (4.3), (4.4) and (4.5) respectively plotted as a function of the distance $\rho$ of the sigmoid from the center of the data, for $\tau = 0.5$.

Figure 2.
A crease projection unit is of the form $\max(X^\prime \theta - \rho, 0)$. The curves $\delta_0$, $\delta_C$ and $\delta_L$ from formulas (4.3), (4.4) and (4.5) respectively plotted as a function of the distance $\rho$ of the crease from the center of the data. The curve $\delta_A$ is an asymptotic value $(\rho^2 + 5)/2$ valid as an approximation for the other $\delta'$s when $\rho$ is large.

Figure 3.
A Cauchy projection unit is of the form $1/2 + \tan^{-1}(\frac{X^\prime \theta - \rho}{\tau})/\pi$. The curves $\delta_C$ from formula (4.4) is plotted as a function of the distance $\rho$ of the crease from the center of the data. A curve is given for $\tau = 4, 2, 1, 0.5, 0.25, 0.125$. They all tend to 1.0 as $\rho$ increases. For $\delta_L$ the limit is 2.0.

Figure 4.
A multiquadric radial basis unit is of the form $(\tau^2 + (X - \rho \theta)(X - \rho \theta))^{1/2}$. The points plotted are $\delta_C$ versus $\rho$ where $\delta_C$ is taken from formula (4.7), with $d = 20, 60, 100$. The points use $\tau = 0$, but with $\tau$ as large as 4.0, the points still overstrike the $\tau = 0$ points. The curves plotted are an asymptotic approximation: $(1 + d/2\rho^2)^{-1}$, accurate for large $\rho$. Their accuracy also appears to be better for larger $d$. 

11
Delta for Logistic Sigmoids
with Tau = 0.5

Delta

Radius rho
Delta for Crease Functions

Delta

Radius rho
Delta_C for Cauchy Sigmoids
Tau from 0.125 to 4
Delta_C for Multiquadric R. B. F. Curves From Asymptotic Formula