Igor Vajda; Jiří Grim
About the maximum information and maximum likelihood principles

*Kybernetika*, Vol. 34 (1998), No. 4, [485]--494

Persistent URL: [http://dml.cz/dmlcz/135236](http://dml.cz/dmlcz/135236)

**Terms of use:**


Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these Terms of use.

This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project DML-CZ: *The Czech Digital Mathematics Library* [http://project.dml.cz](http://project.dml.cz)
ABOUT THE MAXIMUM INFORMATION
AND MAXIMUM LIKELIHOOD PRINCIPLES
IN NEURAL NETWORKS

IGOR VAJDA AND JIŘÍ GRIM

Neural networks with radial basis functions are considered, and the Shannon information
in their output concerning input. The role of information-preserving input transformations
is discussed when the network is specified by the maximum information principle and by
the maximum likelihood principle. A transformation is found which simplifies the input
structure in the sense that it minimizes the entropy in the class of all information-preserving
transformations. Such transformation need not be unique - under some assumptions it may
be any minimal sufficient statistics.

1. INTRODUCTION

In this paper the attention is restricted to the important class of so-called radial
basis function neural networks, which are intensively studied in the recent literature.
These networks were introduced by Bromhead and Lowe (4). Other contributions
can be found in Specht [25], Moody and Darken [19], Lowe [18], Casdagli [5], Poggio
and Girosi [23], Xu et al [33], Streit and Luginbuhl [27], Watanabe and Fukumizu
[31], Ukrainec and Haykin [29] and others. A systematic treatment can be found in
Chap. 7 of Haykin [11] and Chap. 30 of Devroye et al [8].

A radial basis function network (RBF network) consists of several layers. The
input layer is a collection of $d$ real data sources

$$x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d.$$  

The second hidden layer consists of $M$ units. The $m$th unit responds by

$$\phi_m(x) = \frac{1}{\sigma_m} K \left( \frac{x - t_m}{\sigma_m} \right), \quad 1 \leq m \leq M$$

to the input $x$. Here $K : \mathbb{R}^d \to \mathbb{R}$ is a probability density on $\mathbb{R}^d$, symmetric about
$0 \in \mathbb{R}^d$ and called a symmetrical kernel, i.e.

$$K(y) = \varphi(||y||^2) \quad \text{for all } y \in \mathbb{R}^d, \quad \varphi : \mathbb{R} \to \mathbb{R},$$

$^1$The present work has been supported by the Grant of the Academy of Sciences of the Czech
Republic No. A2075703 and partially by the Complex Research Project of the Academy of Sciences
of the Czech Republic No. K1075601.
$t_m \in \mathbb{R}^d$ is a center of symmetry of $\phi_m$, and $\sigma_m > 0$ characterizes a dispersion of responses around this center. All hidden layer units are supposed to be activated by the identity function with a zero threshold, i.e., $(\phi_1(x), \ldots, \phi_M(x))$ is the output of the hidden layer. The output layer consists of $K$ linear units (neurons) responding by

$$\rho_k = \sum_{m=1}^{M} w_{km} \phi_m, \quad 1 \leq k \leq K,$$

(1.1)

to the output $(\phi_1, \ldots, \phi_M) = (\phi_1(x), \ldots, \phi_M(x))$ of previous layer. The network output $y$ is either the vector $(\rho_1, \ldots, \rho_K)$ a deterministic- or stochastic one.

The RBF networks are of a great practical interest because they can easily be realized. Indeed, the $m$th hidden unit can formally be decomposed into a simple linear network consisting of one layer of $d$ nodes with scalar inputs $x_j$, $1 \leq j \leq d$, and constant weights $1/\sigma_m$, using the identity activation functions and thresholds $t_{mj}/\sqrt{\sigma_m}$. The output neuron has weights $w_j$ coinciding with the outputs $(x_j - t_{mj})/\sigma_m$ of the respective nodes, and an activation function equal to the above considered $\varphi$ with a zero threshold.

Possible example of $\varphi$ is the continuous sigmoidal function

$$\varphi(r) = \begin{cases} (2\pi)^{-\frac{d}{2}} e^{-r/2} & \text{if } r \geq 0, \\ (2\pi)^{-\frac{d}{2}} (1 + e^{r/2}) & \text{otherwise}, \end{cases}$$

leading to the Gaussian symmetric kernel

$$\mathcal{K}(y) = (2\pi)^{-\frac{d}{2}} e^{-\|y\|^2/2}.$$  

The hidden nonlinear layer can thus be replaced by two linear layers, so that the whole RBF network can be realized by a three-layer perceptron.

Similar three-layer perceptron realization (see Streit and Luginbuhl [27]) applies also to the more complicated networks with anisotropic RBF's. These differ from the above considered isotropic RBF's by that the argument $(x - t_m)\sigma_m^{-1}$ of $\mathcal{K}(\cdot)$ in the definition of $\phi_m(x)$ is replaced by $(x - t_m)B_m^{-1}$, i.e. that a regular $d \times d$ norm weighting matrix $B_m$ stands at the place of "norm weighting scalar" $\sigma_m$ (cf. pp. 258–259 in Haykin [11]). Then

$$\phi_j(x) = \frac{1}{|\det B_j|} \mathcal{K}((x - t_j)B_j^{-1}).$$

Thus, in particular, the Gaussian kernel leads to multivariate Gaussian probability densities with means $t_j$ and covariance matrices $\Sigma_j = B_j^T B_j$.

Early information-theoretic analyses of perceptual system have been published soon after Shannon [26]. E.g. Attneave [2] analyzed visual perception on the basis of Shannon information, Uttley [30] suggested a network for adaptive pattern recognition using the same information, and many other similar thoughts may be found in various sources.

More recently Linsker [16,17] proposed a learning method based on the principle of maximum information preservation (the infomax principle). This principle
consists in the maximization of the average mutual information between input and output \( x \) and \( y \) of the neural network.

The concept of mutual information has been used also by other authors. Thus Plumbley and Fallside [22] formulated the maximum information preservation principle of Linsker as a minimization of information loss. They assumed the presence of additive Gaussian noise and analyzed a single-layer network to perform the dimensionality reduction. The information loss of their scheme is upper-bounded by the entropy of the reconstruction error and, in this way, the information loss limitation problem is related to the principal component analysis. Some implications of both information principles for neural network learning algorithms has been later analyzed in more details by Plumbley [21].

Atick and Redlich [1] have investigated the principle of minimum redundancy that applies to noisy channels. A linear matrix operator is optimized to minimize a specially introduced redundancy measure. Haykin [11] has shown that, despite the differences, the principle of minimum redundancy and the principle of maximum information preservation lead to similar results.

Kay [14] considered a neural network with input vector divided into primary and contextual part. Again the relationship between the primary- and the contextual subvectors is measured by the average mutual information to analyze the underlying structural dependences.

Becker and Hinton [3] have extended the idea of maximizing mutual information to unsupervised processing of the image of a natural scene. Specifically, their unsupervised learning procedure maximizes the mutual information between higher-level outputs with adjacent receptive fields. Inspired by the work of Becker and Hinton [3], Ukrainec and Haykin [29] developed an information-theoretic model for the enhancement of radar images. For more details about information theoretic approaches to neural networks see chapter 11 in Haykin [11].

The RBF networks are usually optimized by a hybrid method (cf. Hertz et al [12]) which means that only the weights of the third layer are trained in a supervised way whereas the training of the hidden-layer components is unsupervised. This approach has a good reason since consistent estimation of all network parameters is a difficult problem. Unfortunately the global performance of the RBF neural networks strongly depends on the quality of RBF. In particular the information loss caused by improper RBF cannot be repaired by optimizing the weights of the third layer (cf. Grim [10]).

For appropriately specified weights \( w_{km} \) the responses \( \rho_k(x) \) are becoming mixtures of probability densities \( \phi_1(x), \ldots, \phi_M(x) \) which can be viewed as approximations to given data generating probability densities \( f_k(x), 1 \leq k \leq K \).

This opens the possibility to optimize the choice of RBF’s by means of the maximum likelihood principle (ML principle). Of particular interest are iterative statistical schemes leading to maximum likelihood estimates of parameters of RBF’s from given parametrized families. These schemes provide the possibility of iterative learning.

Since the late sixtieth there is an iterative computational scheme called EM algorithm (cf. Dempster et al [7]) which is widely applicable to estimation of mixtures. Design of RBF networks by means of EM algorithm has been studied e. g. by Jacobs
and Jordan [13], Xu and Jordan [32], Haykin [11], Palm [20], Streit and Luginbuhl [27] and Watanabe and Fukumizu [31]).

For the sake of completeness, let us mention that there exist also non-parametric statistical principles leading asymptotically to optimal RBF networks (see Devroye et al [9], in particular Chapter 30, and further references there in, and also Vajda and Grim [28]).

In this paper we are interested in the infomax and ML principles. A difficulty with their application arises when the dimensionality of the input grows. The complexity of application of both the gradient ascent on the input-output Shannon information in the case of infomax, and the EM algorithm in the case of ML, grows with the dimension $d$ of the input space $\mathcal{R}^d$. By using the idea of statistical sufficiency, we characterize a class of transformations $T$ of the input $x$ which preserves the input-output Shannon information $I(x;y)$, i.e. satisfies the relation

$$I(T(x);y) = I(x;y),$$

and has an entropy $H(T(x))$ not greater than $H(x)$ and also not greater then $H(U(x))$ for any transformation $U$ with $I(U(x);y) = I(x;y)$. Since the minimal entropy means a minimal source complexity (in the sense of numerical description, see Risannen [24]) the class of transformations $T$ is an important instrument for reduction of complexity of RBF neural networks.

2. THE RESULTS

Let the weights $w_{km}$ of the above introduced RBF network be nonnegative, and let the network output $Y$ be random in the sense that an output neuron $Y = k$ is selected to be fired with conditional probability

$$Pr(Y = k|x) = \frac{\rho_k(x)}{\sum_{j=1}^{K} \rho_j(x)} \triangleq p_k(x), \quad 1 \leq k \leq K. \quad (2.1)$$

Further let $X$ be a random input distributed by a probability density $f(x)$ on $\mathcal{R}^d$. Then

$$p(x) = (p_1(x), \ldots, p_K(x)) \quad (2.2)$$

is the conditional distribution of $Y$ given $X = x$ and

$$p = (p_1, \ldots, p_K) = \int_{\mathcal{R}^d} p(x) f(x) \, dx$$

is an unconditional distribution of $Y$. Thus

$$I(X;Y) = \int_{\mathcal{R}^d} \left[ \sum_{k=1}^{K} p_k(x) \log \frac{p_k(x)}{p_k} \right] f(x) \, dx \quad (2.3)$$

is the Shannon input-output information.

According to the infomax principle, the information (2.3) is to be maximized. Based on this principle, several iterative learning rules for centers $t_m$ of Gaussian
RBFs and for the weight matrices \( W = (w_{km}) \) performing a gradient ascent on the information have been proposed and successively applied in the literature (see e.g. Haykin [11], pp. 459–460).

Let us suppose that the information (2.3) is finite. Denote by \( E \) the expectation with respect to the distribution \( f(x) \) of \( X \) on \( \mathbb{R}^d \), and consider the strictly concave function

\[
h_K(u) = h_K(u_1, \ldots, u_K) = - \sum_{k=1}^{K} u_k \log u_k
\]

in the domain \( u > 0 \) (with \( 0 \log 0 = 0 \)). By (2.3),

\[
I(X;Y) = H(Y) - H(Y|X) = h_K(p) - Eh_K(p(X)).
\]

If \( T: \mathbb{R}^d \rightarrow T \) is a measurable mapping into a space with \( \sigma \)-algebra \( A \) then

\[
I(T(X);Y) = H(Y) - H(Y|T(X)) = h_K(p) - Eh_K(Ep(X)|T^{-1}A)).
\]

As well known, \( I(T(X);Y) \leq I(X;Y) \) so that

\[
Eh_K(Ep(X)|T^{-1}A)) \geq Eh_K(p(X)). \tag{2.4}
\]

Further, by Jensen’s inequality for conditional expectations (cf. (A.16) on p. 208 of Liese and Vajda [15]),

\[
h_K(Ep(X)|T^{-1}A)) \geq h_K(p(X)) \text{ a.s.} \tag{2.5}
\]

and this relation holds with \( \geq \) replaced by \( = \) if and only if

\[
p(X) = Ep(X)|T^{-1}A) \text{ a.s.} \tag{2.6}
\]

Thus

\[
I(T(X);Y) = I(X;Y) \tag{2.7}
\]

implies that the equality in (2.4) takes place, so that (2.5) holds with the sign of equality and, consequently, (2.6) is satisfied. In other words, (2.7) implies \( T^{-1}A \)-measurability of the function \( p(x) \). It follows from here in particular that if \( T_0(x) = p(x) \) is a mapping from \( \mathbb{R}^d \) into the simplex \( T_0 \subset \mathbb{R}^K \) of stochastic \( K \)-vectors with the \( \sigma \)-algebra \( \mathcal{A}_0 \) of Borel subsets then

\[
T_0^{-1} \mathcal{A}_0 \subset T^{-1} \mathcal{A}. \tag{2.8}
\]

Let now \( E \) and \( T \) be defined as above and consider the strictly convex function \( \psi(u) = - \log u \) in the domain \( u > 0 \), naturally extended to \( u = 0 \). If

\[
E\psi(f(x)) < \infty \text{ and } Ef(x) < \infty \tag{2.9}
\]

then we define entropy of \( T(X) \) by the formula

\[
H(T(X)) = E\psi(E(f|T^{-1}A)).
\]
By the above mentioned Jensen’s inequality it holds for every $T$
\[ H(T(X)) \leq H(X) = E\psi(f(X)). \] (2.10)
and $H(T(X)) \geq 1 - Ef(X) > -\infty$ because $\psi(u) \geq 1 - u$. By the same inequality
the inclusion $B_0 \subset B$ implies
\[ \psi(E(f|B_0)) = \psi(E(E(f|B)|B_0)) \leq E(\psi(E(f|B)|B_0)) \text{ a.s.,} \]
which in turn implies the monotonicity relation
\[ E\psi(E(f|B_0)) \leq E\psi(E(f|B)). \]
From here and (2.8) we obtain the following.

**Assertion 1.** If a measurable mapping $T$ satisfies (2.7) and the RBF network
input $X$ satisfies (2.9) then
\[ H(p(X)) \leq H(T(X)). \] (2.11)
If, moreover, there exists a mapping $T_0$ such that $p(X) = T_0(T(X))$ then the equality
takes place in (2.11).

The first statement follows from the fact that (2.7) implies that $B_0 = T_0^{-1}A_0$
is included in $B = T^{-1}A$ and, by the monotonicity mentioned above, $H(p(X)) = E\psi(E(f|B_0))$
at most equals $H(T(X)) = E\psi(E(f|B))$.

The second statement follows from the first one and from the following obvious
generalization of (2.10): if $T$ is as in (2.10) and $\hat{T} : T \rightarrow \hat{T}$ is measurable then
\[ H(\hat{T}(T(X))) \leq H(T(X)). \]

**Example 1.** Let us consider isotropic Gaussian RBF’s centered at $t_m \in \mathbb{R}^d$
with variances $\sigma^2 > 0$ and let the weight matrix $W = (w_{km})$ be stochastic. Then
\[ \rho_k(x) = \sum_{m=1}^{M} w_{km} (2\pi\sigma_m^2)^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \left( x - t_m \right)^2 \right) \]
where $\bar{x}$ and $s_x^2$ are sample mean and variance specified explicitly below and
\[ f(x) = \frac{1}{K} \sum_{k=1}^{K} \rho_k(x). \]

By the first statement, the vector $p_k(x), 1 \leq k \leq K$ achieves the minimal entropy
among all random transforms $T(X)$ preserving the information. Since the bivariate
statistic
\[ T^*(x) = (T_1^*(x), T_2^*(x)) = (\bar{x}, s_x^2) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i, \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \]
is sufficient for the family \((x_k(x) : 1 \leq k \leq K)\), it satisfies (2.7). Thus by the second statement, \(T^*(X) = (X, S^2)\) achieves the minimum entropy too.

Any optimization, iterative adaptation etc. can thus be based on collections of bivariate data \((X, S^2)\) instead of the \(d\)-variate data \(X\).

The model of Example 1 can be generalized as follows. Let \(p = (p_1, \ldots, p_K)\) be a stochastic \(K\)-vector, \(C = (c_{km})\) a stochastic \(K \times M\) matrix, and let the RBF network weights be defined by

\[
w_{km} = p_k c_{km}.
\]

Then the formula

\[
f_k(x) = \sum_{m=1}^{M} c_{km} \phi_m(x).
\]

defines a family \(F = (f_k : 1 \leq k \leq K)\) of probability densities on \(R^d\). In this case

\[
\rho_k(x) = p_k f_k(x).
\]

Let the density of network input \(X\) be given by the formula

\[
f(x) = \sum_{k=1}^{K} \rho_k(x) = \sum_{k=1}^{K} p_k f_k(x).
\]

The conditional distribution of the network output \(Y\) given \(X = x\) then satisfies the relation

\[
p_k(x) = \frac{p_k f_k(x)}{f(x)}.
\]

and the unconditional distribution is given by the \(K\)-vector \(p\).

The distributions \((p, F)\) define a Bayesian statistical experiment described by a random parameter \(\Theta\) distributed by \(p\) and a random observation \(X\) conditionally distributed by (2.12) given \(\Theta = k\). The pair \((\Theta, X)\) has the same distribution as \((X, Y)\). Therefore

\[
I(X; \Theta) = I(X; Y).
\]

If \(\Theta\) is the input of channel \(C\) then the output is the random variable \(Z\) with

\[
Pr(Z = m) = \sum_{k=1}^{K} p_k c_{km} \Delta = q_m.
\]

All random variables under consideration form a Markov chain \(\Theta \rightarrow Z \rightarrow X \rightarrow Y\).

It follows from here (cf. Theorem 2.8.1 in Cover and Thomas [6])

\[
I(\Theta; X) \leq I(\Theta, Z; X) = I(Z; X).
\]

For network inputs conditionally distributed by mixtures (2.12) with anisotropic Gaussian RBF's \(\phi_m(x)\), Grim [10] studied the ML estimator of weights \(c_{km}\) and parameters implicitly figuring in functions \(\phi_m(x)\). He established the convergence
of EM algorithm leading to iterative specification of the network, based on independent samples of data $X_{k1}, \ldots, X_{kn}$ distributed by $f_k$ for $1 \leq k \leq K$. In this context an important role plays the descriptive Bayesian experiment $(q = (q_1, \ldots, q_M), \mathcal{F}_0 = (\phi_m : 1 \leq m \leq M))$, with the unknown parameter $Z$, sample $X$ distributed conditionally under $Z = m$ by $\phi_m$ and unconditionally by

$$\sum_{m=1}^{M} q_m \phi_m(x) = f(x),$$

where $f(x)$ is given by (2.13): In (2.14), $I(Z; X)$ is the upper bound on the information $I(\Theta; X)$ concerning the inference parameter $\Theta$. This shows that the quality of RBF’s in the family $\mathcal{F}_0$ is limiting any further decision making, i.e. that a possible information loss caused by inaccurate components $\phi_m$ cannot be repaired by optimizing the weights $w_{km}$.

Preservation of the information $I(Z; X)$ during manipulations with data, and at the same time, the need to simplify the data structure, underline the importance of the assertion that follows. In this from the formal point of view special version of Assertion 1 we consider the conditional distribution $q(x) = (q_1(x), \ldots, q_M(x))$ of $Z$ given $X = x$, given by

$$q_m(x) = \frac{q_m \phi_m(x)}{f(x)}.$$

Note that the relation

$$I(Z; X) = I(Z; T(X))$$

holds for every statistic $T : \mathbb{R}^d \rightarrow T$ which is sufficient for $\mathcal{F}_0$. Similarly it follows from (2.12) and (2.13), and from the convexity of logarithmic function $\varphi(u)$ and of quadratic function $\psi(u) = u^2$ figuring in (2.9), that $X$ satisfies (2.9) if

$$- \int_{\mathbb{R}^d} \phi \hat{m}(x) \log \phi_m(x) \, dx < \infty \quad \text{and} \quad \int_{\mathbb{R}^d} \phi_m(x)^2 \, dx < \infty$$

for all $1 \leq m, \hat{m} \leq M$.

**Assertion 2.** If a measurable mapping $T$ satisfies (2.15) and the RBF’s satisfy (2.16) then

$$H(q(X)) \leq H(T(X)).$$

If, moreover, there exists a mapping $T_0$ such that $q(X) = T_0(T(X))$ then the equality takes place in (2.17).

**Example 2.** Let $\mathcal{F}_0$ be a family of exponential densities

$$\phi_m(x) = h(x) c_m \exp \left( \sum_{j=1}^{J} w_{mj} \tau_j(x) \right)$$

satisfying (2.16). Since the statistic $\tau(x) = (\tau_1(x), \ldots, \tau_J(x))$ is sufficient for $\mathcal{F}_0$, it minimizes the entropy in the class of transformations $T(x)$ satisfying (2.15).

(Received December 18, 1997.)
REFERENCES


e-mails: vajda@utia.cas.cz, grim@utia.cas.cz