General Gaussian Priors for Improved

Generalization

D Saad *

Department of Physics, The University of Edinburgh,

Edinburgh EH9 3JZ, U.K. †

Abstract

We explore the dependence of performance measures, such as the generalization error and generalization consistency, on the structure and the parameterization of the prior on ‘rules’, instanced here by the noisy linear perceptron. Using a statistical mechanics framework, we show how one may assign values to the parameters of a model for a ‘rule’ on the basis of data instancing the rule. Information about the data, such as input distribution, noise distribution and other ‘rule’ characteristics may be embedded in the form of general gaussian priors for improving net performance. We

*Acknowledgements: The author would like to thank Peter Sollich and Alastair Bruce for reading the manuscript as well as for useful suggestions, and Alan Dunmur for useful discussions.
†Requests for reprints should be sent to David Saad, Department of Computer Science & Applied Mathematics, Aston University, Birmingham, B4 7ET, U.K.
examine explicitly two types of general gaussian priors which are useful in some simple cases. We calculate the optimal values for the parameters of these priors and show their effect in modifying the most probable, MAP, values for the rules.

**Keywords:** Learning and generalization, Regularizers, Priors.

1 **Introduction**

Modelling an unknown rule from a set of input-output relations requires certain assumptions or hypotheses, prior to the training procedure itself. If there is no prior knowledge assumed, the model is likely to fail (Wolpert, 1992), i.e., to show poor resemblance to the ‘true rule’ instanced by the data. The reason for this is the infinite number of models that can potentially generate any given data set. Prior convictions (hypotheses) about the ‘true rule’ are expressed, *de facto*, in the form chosen for the model (network configuration), and in the values assigned to a further set of parameters, characterizing constraints on the set of possible models. The role of these constraints is to reduce the volume of the relevant parameter space or in other words to reduce its effective VC dimension (Solla, 1992), directing the solutions towards the region in which the ‘true rule’ is likely to be found. The simplest form of these constraints is the weight decay (Weigend, Huberman and Rumelhart, 1990) used to enhance generalization capabilities and to prune redundant parameters.
In general terms the modelling problem is as follows. We are given a set of data $D$ comprising $p$ members; each member instances an unknown rule (or mapping) $R^0$ connecting the components of the member (the ‘input’ and the ‘output’). We attempt to model $R^0$ by some rule $R(\{w\})$ specified by a parameter set $\{w\}$ having $N$ members. The set of parameters characterizing constraints on the set $\{w\}$ will be referred to as $\{\beta\}$.

Many techniques have been suggested for imposing restrictions on the space of possible models and reducing the number of free parameters involved (Weigend et al, 1990; LeCun, Denker and Solla, 1990; Hassibi and Stork, 1993) and were shown to be useful for improving network performance. Some of these methods have also been examined analytically for simple cases (Solla, 1992; Dunmur and Wallace, 1993; Bruce and Saad, 1994; Marion and Saad, 1995). Here, we focus on the assignment of the hypothesis parameters $\{\beta\}$ in the framework of general gaussian distributions for the prior parameters and for the input data. The generalization of the isotropic gaussian prior enables us to embed additional information in the prior on the rules, for optimizing network performance. Practically, the new prior may appear as a penalty term, added to the training error, modifying the training procedure by a new weight decay term (since stochastic learning with weight decay is generally equivalent to having a prior on the rules (Seung, Sompolinsky and Tishby, 1992)).

The analysis is carried out in a context of sufficiently large $N$ and $p$ and for a stochastic training algorithm, where the methods of statistical mechanics are
appropriate, and which is relevant to neural network modelling (for general review see Hertz, Krogh and Palmer, 1991). In this framework, both the number of examples, \( p \to \infty \), and the network size, \( N \to \infty \), while keeping their ratio \( \alpha = p/N \) finite. The techniques and the terminology used in this work rely heavily on the works of Krogh and Hertz (1992), MacKay (1992a,1992b) and Bruce et al (1994).

Using methods adopted from statistical mechanics, Krogh et al (1992) have explored the particular case in which the rule connecting input and output is linear, the data is corrupted with gaussian noise, and the parameters \( \{w\} \) of the model rule are drawn from a gaussian prior. For this case (the ‘noisy linear perceptron with weight decay’, to be referred to as NLP) they identify a linear dependence between the two hypothesis parameters used in their work, characterizing assignments that are ‘optimal’ in the sense that they minimize the typical value of the error with which the model will predict the output associated with a new input (the generalization error).

A different approach, based on a Bayesian framework, for assigning the hypothesis parameters \( \{\beta\} \) was developed and explored by MacKay (1992a,1992b). In this framework the hypothesis parameters \( \{\beta\} \) are assigned values such as to maximize the conditional probability \( P(D|\{\beta\}) \) – the ‘evidence’ – for the parameters, which in the absence of prior information on the hypothesis parameters may provide a direct measure of the conditional probability \( P(\{\beta\}|D) \). The parameters which maximize the evidence are thus ‘optimal’ in the sense that they represent the
single most likely (maximum a posteriori, MAP) values given the data. For the linear perceptron, these optimal parameters can be shown to be identical to those which optimize the generalization error and generalization consistency (Bruce et al, 1994).

The main contribution of the evidence procedure is for real world problems where no information about the ‘true rule’ is available and where probabilities, required for defining evidence and the optimal hypothesis parameters \( \{ \beta \} \), can be calculated explicitly on the basis of the data. In this work, in the absence of real data, we will assume prior knowledge about the ‘true rule’ and use the statistical mechanics framework to average the performance measures over ‘quenched’ data variables. We will concentrate on priors which optimize the performance measures directly, assuming, on the basis of Bruce et al (1994), similarity between our results and those obtained via the evidence procedure.

The main incentive for introducing general gaussian priors is that they enable us to embed more information about the ‘true rule’, the distribution of inputs, noise sources etc., as constraints, reducing the space of solutions and improving our generalization capabilities. This is likely to be the case in many real world problems, where one can expect a non-uniform noise and input distribution, especially since various components of the input data may come from different sources. In these cases, additional information is readily accessible and may be used to improve constraints and network performance.

However, an effective use of general gaussian priors, as well as other more compli-
cated priors, requires more information. In the absence of additional information one should stick to simple priors, since the complexity of the prior should reflect the amount of available information and might actually damage the performance of the model otherwise.

One should also mention that general gaussian priors as well as other types of regularizers are commonly used for linear additive models (Hastie and Tibshirani, 1990). More complicated priors may be used for special cases, however, it is rather complicated to analyze them theoretically. General gaussian priors have been previously suggested and used by MacKay (1992b), although their effect on optimizing net performance has not been analyzed.

In the next section we will define and calculate the performance measures in the case of a general gaussian input distribution as well as a general gaussian prior. In sections 3 and 4 we will examine two special cases of gaussian priors, find the optimal set of parameters, and show the behaviour of the performance measures for optimal and sub-optimal values. In section 5 we summarize the main results.

2 Calculating performance measures

We suppose that our data $D$ comprises $p$ input-output pairs $(x^\mu, y^\mu)$ where the vector $x^\mu$ has elements $x_j^\mu (j = 1 \ldots N)$. Although the problem can be formulated in general terms for any student and teacher configurations, the calculation can be carried out explicitly only for linear nets. We therefore suppose that the
input-output relation $\mathcal{R}^0$ is linear, and subject to corruption by a gaussian noise:

$$y^\mu = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} w_j^0 x_j^\mu + \eta \tag{2.1}$$

where $w_j^0 (j = 1 \ldots N)$ are the elements of a rule vector $w^0$, and $\eta$ is a random gaussian noise variable of variance $\sigma^2$. We suppose that the elements $x_j^\mu$ of the inputs are gaussian random variables derived from a general gaussian distribution

$$P(x) = (2\pi)^{-N/2} \Xi^{-1/2} e^{-\frac{1}{2} x^T \Xi^{-1} x} \tag{2.2}$$

where $\Xi$ is assumed to be a positive definite matrix.

We proceed to examine the effectiveness of ‘models’ $\mathcal{R}$ of the data-generating process which takes the rule to be linear, parameterized by a rule vector $w$,

$$\mathcal{R}(x^\mu) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} w_j x_j^\mu \tag{2.3}$$

with elements $\{w\}$ which take the corruption process to be gaussian with variance $1/2\beta$ and are subjected to a general gaussian prior of the form

$$P(w|\mathcal{B}, \beta) = \pi^{-N/2} |\mathcal{B}|^{1/2} e^{-w^T \beta \mathcal{B} w} \tag{2.4}$$

representing the hypothesis about the rule itself through the conditional probability (prior) $P(w|\mathcal{B}, \beta)$, where $\mathcal{B}$ is assumed to be a positive definite matrix.
The quantities $\beta$ and $\mathcal{B}$ constitute the ‘hypothesis parameters’. Note that here we generalized the conventional isotropic gaussian prior, including many more free parameters, which should typically be defined using additional information about the ‘true rule’ and the input distribution.

The hypothesis about the corruption process is expressed through the further conditional probability (likelihood):

$$P(D|\mathbf{w}, \beta) = \left(\frac{\beta}{\pi}\right)^\frac{p}{2} \prod_{\mu=1}^p e^{-\beta (\nu^\mu - \frac{1}{\sqrt{p}} \sum_{j=1}^N w_j x_j^\mu)^2}$$ (2.5)

Appealing to Bayes Theorem (Papoulis, 1986), these two conditional probabilities may be combined to give a conditional probability (the posterior) for the model rule, parametrized by the rule vector $\mathbf{w}$:

$$P(\mathbf{w}|D, \mathcal{B}, \beta) = \frac{P(D|\mathbf{w}, \beta)P(\mathbf{w}|\mathcal{B}, \beta)}{P(D|\mathcal{B}, \beta)}$$ (2.6)

The normalizing factor in this relation, $P(D|\mathcal{B}, \beta) = \sum_{\mathbf{w}} P(D|\mathbf{w}, \beta)P(\mathbf{w}|\mathcal{B}, \beta)$ constitutes the ‘evidence’ which the data provides for the assignment of the model parameters in a method introduced by MacKay (1992a,1992b). For the NLP, and in the thermodynamic limit, the optimal parameters obtained by maximizing the evidence are equal to those obtained by minimizing the generalization error and the generalization consistency, defined below. Therefore, we concentrate in this work on the optimal parameters with respect to the generalization error.
and generalization consistency, using the posterior probability for calculating the performance measures.

The are many forms of performance measures, estimating the resemblance between the ‘true rule’ and the model (Krogh et al, 1992; Levin, Tishby and Solla, 1990; Hansen, 1993). The most widely used performance measure is the generalization error $\epsilon_g$ defined by

$$\sigma^2 \epsilon_g \equiv \left\langle \left[ \langle R^0(x) - R(x) \rangle_w \rangle_x \right]^2 \right\rangle_{xy}, \quad (2.7)$$

measuring the average squared error, over test vectors, between the average student rule and the non-corrupted teacher rule. The coefficient $\sigma^2$, representing the noise level in the training set, appears in all of the terms and is therefore factored out. Other performance measures may examine the average squared error between the average student rule and the corrupted teacher rule or the average squared error, over both test vectors and rules, between all student rules and the teacher.

The inner average $\langle \cdot \rangle_w$ extends over the ensemble of model rules for given data; the average $\langle \cdot \rangle_x$ extends over the input data distribution (for the test data); the outer average $\langle \cdot \rangle_{xy}$ extends over the quenched variables (all possible training data and the noise).

We will now carry out the calculation explicitly for linear rules (Eq. 2.1). Using the posterior probability (Eq. 2.6) and on the basis of Eqs.(2.2) and (2.4) we can
carry out the first two averages straightforwardly (for details see appendix A), obtaining:

$$
\epsilon_2 = \langle \frac{1}{N} \text{Tr} (A + L)^{-1} \rangle - \frac{1}{N} \sum_{j,k} \left( B - \frac{1}{\sigma^2} B \mathbf{w}_\mu^\top \mathbf{w}_\mu B \right)_{jk} \left( (A + L) \Xi (A + L) \right)_{jjk}^{-1}
$$

where

$$
A \equiv \frac{1}{N} \sum_{\mu} \Xi^{-1} x_\mu x_\mu^\top
$$

$$
L \equiv \Xi^{-1} B
$$

The unusual form of the input correlation matrix $A$ and the prior term $L$, is due to the anisotropic input distribution.

Calculating the quenched average, over different data sets, is a difficult task. Since the only terms that depend on the training data are of the form $(A + L)^{-1}$ and $(A + L)^{-1} \Xi^{-1} (A + L)^{-1}$, the averages may be carried out by adapting the analysis of Hertz, Krogh and Thorbergsson (1989) or by another method, recently suggested by Sollich (1994). The result of the quenched average of $(A + L)^{-1}$, termed the response function, $\mathcal{G}$, is a solution of a matrix equation of the form:

$$
\mathcal{G} \equiv \langle (A + L)^{-1} \rangle = \left( L + \frac{\alpha}{1 + \frac{1}{N} \text{Tr} \mathcal{G}} \right)^{-1}
$$
where \( \alpha \equiv p/N \). In particular, one may consider the trace of \( G \),

\[
G \equiv \frac{1}{N} \text{Tr} \ll (A + L)^{-1} \gg = \frac{1}{N} \text{Tr} \left( L + \frac{\alpha}{1 + G} \right)^{-1}
\]  

(2.11)

using the identity:

\[
\ll \text{Tr} \ C(A + L)^{-1} \gg = \text{Tr} \ C \left( L + \frac{\alpha}{1 + G} \right)^{-1}
\]  

(2.12)

where \( C \) is a general matrix and \( G \) is the solution of Eq. (2.11), which is sufficient for calculating the averages. Using equation (2.11) and the response function \( G \) one can derive the explicit expression for the generalization error:

\[
\epsilon_g = G + \sum_{jk} \left( B - \frac{1}{\sigma^2} w^j w^T B \right) \frac{\partial G}{\partial z_{jk}}
\]  

(2.13)

where \( \varepsilon_{ij} \to 0 \) are elements added to the matrix \( \Xi(L + A) \). The infinitesimal elements \( \varepsilon_{ij} \) are used in conjunction with the identity:

\[
[(L + A)^{-1} \Xi^{-1} (L + A)^{-1}]_{ij} \equiv - \lim_{\varepsilon \to 0} \frac{\partial}{\partial \varepsilon_{ij}} \text{Tr} (L + A)^{-1}
\]  

(2.14)

for calculating averages such as \( \ll (L + A)^{-1} \Xi^{-1} (L + A)^{-1} \gg \).

This general expression (Eq.2.13) reduces to the conventional form (Krogh et al., 1992, Bruce et al., 1994) of \( \epsilon_g \) when the general input gaussian distribution is replaced by an isotropic gaussian distributions with variance \( \sigma_x^2 \) and a scalar
prior coefficient $\lambda$ replaces the matrix $B$.

Another measure for the effectiveness of the modelling procedure is the generalization consistency $\delta_g$ measuring the variance of the model output for given fixed data and test input:

$$\sigma^2 \delta_g \equiv \ll [R(x) - \langle R(x) \rangle_w]^2 \rangle_w,x \gg -\sigma^2 \epsilon_g$$  \hspace{1cm} (2.15)$$

A small value of $\delta_g$ implies that the variance of the model output corresponds to the average generalization error, i.e., that the model predicts its own errors accurately. In the present context we find that

$$\delta_g = \frac{1}{2^3 \sigma^2} G - \epsilon_g$$  \hspace{1cm} (2.16)$$

3 General diagonal gaussian priors

We will now turn to a specific example of a general gaussian prior which is the natural extension of the isotropic gaussian prior – the anisotropic diagonal gaussian prior.

3.1 Generalization error

For simplicity we will examine the case in which the noise and input distributions are simple gaussian distributions with variances $\sigma^2, \sigma_x^2$ respectively, and the
penalty matrix has only two different values on the diagonal. The case where
the input distribution is also represented by an anisotropic diagonal gaussian is
a natural extension of our results and can be easily derived.

In this case the general form of the generalization error (Eq. 2.13) can be reduced to:

\[
\epsilon_g = G + \left[ \frac{\lambda_1 (\kappa - \lambda_1 \nu_1)}{(\lambda_1 + \Omega)^2} + \frac{\lambda_2 (1 - \kappa - \lambda_2 \nu_2)}{(\lambda_2 + \Omega)^2} \right] \left[ \frac{\kappa \Omega^2}{\alpha (\lambda_1 + \Omega)^2} + \frac{(1 - \kappa) \Omega^2}{\alpha (\lambda_2 + \Omega)^2} - 1 \right]^{-1}
\]  

(3.17)

where:

\[
\begin{align*}
\lambda_1 &= \mathcal{B}_{ii}/\sigma_x^2 \quad \text{for } 0 < i \leq k \\
\lambda_2 &= \mathcal{B}_{ii}/\sigma_x^2 \quad \text{for } k < i \leq N \\
\nu_1 &= \frac{\sigma^2 \sigma_{w1}^2}{\sigma^2} \\
\nu_2 &= \frac{\sigma^2 \sigma_{w2}^2}{\sigma^2} \\
\sigma_{w1}^2 &= \frac{1}{N} \sum_{i=1}^{k} w_i^2 \\
\sigma_{w2}^2 &= \frac{1}{N} \sum_{i=k+1}^{N} w_i^2 \\
\kappa &= k/N \\
\Omega &= \frac{\alpha}{1 + G}
\end{align*}
\]

and \( G \) is the solution of Eq.(2.11) for this particular case where:

\[
L_{ij} = \delta_{ij} \begin{cases} 
\lambda_1 & \text{for } 0 < i \leq k \\
\lambda_2 & \text{for } k + 1 < i \leq N
\end{cases}
\]

(3.18)

An explicit expression for \( G \), the solution of Eq.(2.11) for this case, can be found in appendix B.
3.2 Optimal weight decay

Taking partial derivatives of the generalization error (Eq. 3.17) with respect to \( \lambda_1 \) and \( \lambda_2 \), it is easy to verify that the optimal value for the generalization error is obtained when \( \lambda_1 = \kappa / \nu_1 \) and \( \lambda_2 = (1 - \kappa) / \nu_2 \), i.e., when each one of the penalty elements is set to match the average noise-to-signal ratio related to the corresponding elements of the weight matrix. This result is consistent with the optimal parameter obtained for the isotropic prior case \( \lambda = 1 / \nu \). The prescription for obtaining the optimal values can be extrapolated to the general diagonal penalty matrix in which \( \lambda_i^{opt} = \kappa_i / \nu_i \), where \( \nu_i = \frac{\kappa^2}{2} \frac{1}{N} \sum_{l=1}^{k_i} w_i^2 \) for each group of weight components \( i \). These optimal values are in agreement with the values obtained numerically as well as with those obtained analytically for the large \( \alpha \) expansion. Figure 1 provides a simple example: it shows \( \epsilon_g \) as a function of the number of patterns \( \alpha \) for several values of \( \lambda_1 \), where \( \lambda_2 \) is fixed to its optimal value. We used the following values for the parameters: \( \kappa = 1 / 2, \nu_1 = 0.1 \) and \( \nu_2 = 1 \). It is easy to see that the minimal value for \( \epsilon_g \) is achieved for the optimal values calculated above, whereas typical behavior for undervalued and overvalued priors, with respect to the optimal parameters, can be seen for the extreme cases (for example underestimated noise to signal ratio creates a hump where \( \alpha \) equals the number of effective free parameters). Figure 2 provides an example for the behaviour of \( \epsilon_g \) (as a function of the number of patterns \( \alpha \)) for various choices of \( \kappa \), for fixed values of the matrix parameters \( \lambda_1 = 5 \) and \( \lambda_2 = 0.1 \) and similar val-
ues for \( \nu_1 \) and \( \nu_2 \) as before. In this example one of the parameters, \( \nu_1 \), is set close to its optimal value (slightly undervalued or overvalued depending on the value of \( \kappa \)), while the other is significantly undervalued for most of the values of \( \kappa \). It is easy to see that the characteristic behaviour of \( \epsilon_g \) for undervalued parameters, i.e., a hump where the number of examples equals the effective number of free parameters, becomes dominant as \( \kappa \) decreases. Obviously, better performance is achieved when more weight components have higher signal to noise ratio, i.e., \( \kappa \) is larger in this case. The main difference from the conventional case of scalar gaussian prior is the interplay between the two (possibly different) behaviours of the two sets of weights and priors.

The optimal value of \( \beta \) is derived by optimizing \( \delta_g \) since \( \epsilon_g \) is independent of \( \beta \) for fixed values of \( B \), giving an optimal value of \( \beta = 1/2\sigma^2 \) as expected.

### 3.3 Stability analysis for the isotropic weight decay

In order to justify the use of a more complicated prior one can examine the stability of the isotropic gaussian prior; i.e., given an optimal scalar diagonal weight decay \( \lambda \), under what conditions will an infinitesimal modification \( \delta \) in \( k \) of its components decrease the value of the generalization error.

Expanding \( \epsilon_g \) around the optimal value (Bruce et al, 1994) of the weight decay
parameter $\lambda_{opt} = 1/(\nu_1 + \nu_2)$ one obtains for the first order in $\delta$:

$$\frac{\partial \epsilon_2}{\partial \delta} = f(\nu_1, \nu_2, \alpha, \lambda_{opt})(\kappa \nu_2 - (1 - \kappa)\nu_1)$$  \hfill (3.19)

where $f(\nu_1, \nu_2, \alpha, \lambda_{opt})$ is a function which is always positive. The only case in which $\epsilon_2$ is optimal for $\delta \to 0$, is if $\kappa \nu_2 = (1 - \kappa)\nu_1$, i.e., if the signal-to-noise ratio from the two sections is identical. In any other case, the optimal generalization error is obtained for $\lambda_{i_{opt}}^2 = \kappa_i/\nu_i$.

### 3.4 MAP solutions

In order to understand the effect of the general diagonal penalty term on the maximum a posteriori solutions, we expand the training error

$$E = \sum_{\mu} (y^\mu - \frac{1}{\sqrt{N}}w^T \mu x^\mu)^2 + w^T B w$$  \hfill (3.20)

around the minimum solution $w^*$ for the $B \to 0$ case, to obtain the relationship between $w^*$ and the new minimum $w^B$ with a general penalty matrix $B$ (Hastie et al, 1990):

$$w^B = (H + B)^{-1} H w^*$$  \hfill (3.21)

where $H$ is the Hessian matrix for the $B \to 0$ case, measuring the curvature of the energy surface at this point. In the principal axis representation, the Hessian matrix is diagonal with eigenvalues $\rho_i$. If the Hessian matrix and the penalty
matrix $\mathcal{B}$ commute (which is for example the case for the large $\alpha$ limit), it is easy to see the effect of the general diagonal prior, on the MAP solution. In this case Eq. (3.21) can be explicitly written as

$$w_i^g = \frac{w_i^* \beta_i}{\rho_i + \lambda(i)}$$

(3.22)

which is equivalent to an effective reduction of the VC dimension (Solla, 1992) by $\frac{1}{N} \sum_i \frac{\rho_i}{\rho_i + \lambda(i)}$, where $\lambda(i)$ represents the prior parameter related to the $i$-th weight component. Here, each one of the weight components is suppressed according to its eigenvalue and the penalty term characterizing his group of components only, whereas in the isotropic diagonal case, the penalty term is common to all elements.

4 Isotropic diagonal with isotropic off-diagonal elements

This is another possible extension of the isotropic diagonal penalty term, suppressing differences (between components) within groups of weights. In order to use this type of prior effectively one should group weights with similar characteristics, e.g., weights with a distinct average. The form of the penalty matrix in
this case is

$$B = \sigma_x^2 \left( \begin{array}{ccc} A & 0 & \ldots \\ 0 & A & \vdots \\ \vdots & \ddots & \ddots \end{array} \right)$$

(4.23)

where $A_{ik} = \lambda_d \delta_{ik} + \lambda_o$. Both $0$ (a zero matrix), and $A$ are $M \times M$ matrices. This new prior divides the weights effectively to groups of size $M$ where $\lambda_d$ and $\lambda_o$ are the diagonal and off-diagonal elements respectively for the prior related to each group. Note that this is one possible rotation of the two-component diagonal prior discussed in the previous section.

### 4.1 Generalization error

For simplicity we will examine only the case in which the noise and input are derived from simple gaussian distributions with variances $\sigma^2$, $\sigma_x^2$ respectively. Also here, the extension of the analysis and the results to the case of anisotropic diagonal gaussian input distribution is straightforward.

In this case the general form of the generalization error (Eq. 2.13) can be reduced to:

$$
\epsilon_2 = G_d' + \left[ \lambda_d + \lambda_o - \lambda_d^2 \mu_1 - \lambda_o (M \lambda_o + 2 \lambda_d) \mu_2 \right] G_d'' \\
+ \left[ \lambda_d^2 (\mu_1 - \mu_2) + (M - 1) \lambda_o (1 - (M \lambda_o + 2 \lambda_d) \mu_2) \right] G_o'' 
$$

(4.24)
\[
\mu_1 = \frac{\sigma^2_{w_i}}{\sigma^2} \quad \mu_2 = \frac{\sigma^2_{w_{i}}}{{\sigma^2}^{2}} \\
\sigma^2_w = \frac{1}{N} \sum_{i=1}^{N} w_i^2 \quad \sigma^2_{w_{i}} = \frac{1}{N} \sum_{i=1}^{N} \left( \sum_{k=i}^{(i-1)M+1} w_i \right)^2
\]

and \(G_d\) and \(G_o\) are the diagonal and off-diagonal values of the solution of Eq. (2.10) for this particular case; \(G'\) represents the derivative of \(G\) with respect to \(\lambda_d\).

Since all the penalty matrices \(\mathcal{A}\) are along the diagonal, we could in principle, choose a more general framework in which we concatenate a set of \(P\) matrices \(\mathcal{A}_i\) each of a different size \(M_i \times M_i\) and different diagonal and off-diagonal coefficients \(\lambda_{d_i}^i\) and \(\lambda_{o_i}^i\) respectively. Optimization of the various elements will then be an extension of the procedure described below.

### 4.2 Optimal weight decay parameters

Taking the partial derivative of the generalization error (Eq. 4.24) with respect to \(\lambda_d\) when \(\lambda_o\), it is easy to verify that the optimal values of the generalization error are obtained when \(\lambda_d = \frac{M-1}{M_{\mu_1}-\mu_2}\) and \(\lambda_o = \frac{\mu_1-\mu_2}{\mu_2(M_{\mu_1}-\mu_2)}\). In this case the off-diagonal terms attract the weight components towards the average weight component while the diagonal terms suppress redundant components. The interplay between the two terms is optimal when the contribution from the penalty term counterbalances the noise. This equality results in two equations, one for \(\text{Tr} \ B^{-1}\) and one for \(\sum_{ij} B^{-1}_{ij}\) from which the optimal values for \(\mu_1\) and \(\mu_2\) can be obtained. The optimal solution for the two penalty matrix elements reduces to the conventional solution \(\lambda_d = 1/\mu_1\) and \(\lambda_o = 0\) in the case of \(\mu_2 = \mu_1\).
Figure 3 shows $\epsilon_g$ as a function of the number of patterns $\alpha$ for several values of the decay parameter $\lambda_\alpha$ where $\lambda_d$ is fixed to its optimal value. We have chosen the following values for the parameters: $M = 5$, $\mu_1 = 3$ and $\mu_2 = 10$. It is easy to see that the minimal value for $\epsilon_g$ is achieved for the optimal values calculated above, where typical behavior for undervalued and overevaluated priors can be seen for sub-optimal values. The divergence of $\epsilon_g$ (and a hump for sub-critical values) appears where $M\lambda_\alpha + \lambda_d = 0$. A similar behaviour of $\epsilon_g$ occurs when $\lambda_d$ varies and $\lambda_\alpha$ is set to its optimal value.

Here too, the optimal value of $\beta$ is $\beta = 1/2\sigma^2$, obtained by optimizing $\delta_\beta$.

### 4.3 Optimal group size

We can carry out a similar optimization procedure for the group size $M$. Taking the partial derivative of the expression for the generalization error (Eq. 4.24) with respect to $M$, fixing $\lambda_d$ and $\lambda_\alpha$ to their optimal values, we obtain an expression which is always *positive*, i.e., optimal performance is obtained with a minimal group size. However, the minimal value for $M$ is bounded from below by the value $M = \max \{0, -\lambda_d/\lambda_\alpha\}$ which corresponds to zero variance for the optimal values; smaller values of $M$ will cause $\epsilon_g$ to diverge, corresponding to un-physical values for $\mu_1$ and $\mu_2$.

An example for the dependence of $\epsilon_g$ on the number of patterns $\alpha$ for several values of the group size $M$ is shown in Figure 4. We used the following values for
the parameters: \( \mu_1 = 1, \mu_2 = 10 \) and the decay parameter \( \lambda_d \) and \( \lambda_s \) are fixed to their optimal values. The minimal value for \( \epsilon_g \) is achieved for the optimal value \( M = 11 \).

### 4.4 Stability analysis

We will now examine the stability of the isotropic weight decay with respect to small modifications in the off-diagonal terms, i.e., given an optimal scalar diagonal weight decay \( \lambda \), in what conditions will an infinitesimal modification \( \delta \), in concatenated square matrices of size \( M \times M \) along its diagonal, decrease the value of the generalization error.

Expanding \( \epsilon_g \) around the optimal value of the weight decay parameter (\( \lambda^{opt} = 1/\mu_1 \)) one obtains the first order variation with \( \delta \)

\[
\frac{\partial \epsilon_g}{\partial \delta} = f(\mu_1, \mu_2, \alpha, \lambda^{opt})(\mu_1 - \mu_2) \tag{4.25}
\]

where \( f(\mu_1, \mu_2, \alpha, \lambda^{opt}) \) is a function which is always positive. Obviously the diagonal solution is stable if \( \mu_1 = \mu_2 \); in any other case the diagonal solution is generally unstable. We always prefer smaller off-diagonal values if the average is relatively larger than the variance (\( \mu_1 < \mu_2 \)) and larger off-diagonal values otherwise (note that \( \lambda_s \) is usually negative).
4.5 MAP solutions

Using a similar expansion as in the previous case, we can now use Eq. (3.21) to examine the relation between \( \mathbf{w}^g \) and \( \mathbf{w}^* \). If the Hessian matrix \( \mathbf{H} \) and the penalty matrix \( \mathbf{B} \) commute, Eq. (3.21) can be explicitly written for this case

\[
\mathbf{w}_i^g = \mathbf{w}_i^* - \frac{\rho_i}{\rho_i + \lambda_d} \lambda_o \frac{\sum_k \mathbf{w}_k^* \rho_k}{1 + \sum \lambda_o / (\rho_i + \lambda_d)}
\]  

(4.26)

where \( \rho \) are the eigenvalues of the Hessian matrix and \( B_{ij} = \delta_{ij} \lambda_d + \lambda_o \). In order to get a better understanding of the effect of the prior we will consider the case in which \( \lambda_o \ll \lambda_d + \rho_i \). In this case

\[
\mathbf{w}_i^g = \mathbf{w}_i^* - \frac{\rho_i}{\rho_i + \lambda_d} \lambda_o \sum_k \mathbf{w}_k^* \frac{\rho_k}{\rho_k + \lambda_d}
\]  

(4.27)

We see that each weight element is suppressed with respect to the corresponding eigenvalue \( \rho_i \), and shifted towards the weight average by the second term. The average is weighed by the eigenvalues relates to each one of the weight components.

5 Discussion

In this paper we have explored the issue of general gaussian priors in the context of the noisy linear perceptron. We have formulated the expressions for the
generalization error and generalization consistency using general gaussian input distributions and priors.

We analyzed explicitly two cases of general gaussian priors, representing two different extensions to the isotropic diagonal penalty matrix. We have found the optimal solutions for these cases and have shown that the conventional isotropic diagonal form is unstable with respect to small perturbations in the penalty matrix, showing that even for simple cases it is useful to introduce more complicated penalty terms than the conventional isotropic diagonal form, provided that additional data is available.

The two cases examined enable us to deal more effectively with input vectors and rules fragmented to their typical characteristic sub-vectors, e.g., sub-vectors with distinct averages, which are typical to the case where the input data comes from different sources. Further generalization can be obtained by using general gaussian distributions for the input data as demonstrated in section 2.

Obviously, one should have more information in order to be able to use the general penalty matrix effectively, and misleading information might actually damage net performance, as is also the case for any prior. However, some of that information is easily accessible; for example, in the case where the input information comes from different sources one can estimate the noise and input distribution for each one of the sources separately.

The extent to which using general gaussian priors can improve performance of neural networks is problem dependent and requires further study.
References


A Calculating the generalization error

In order to simplify the derivation we will first calculate the expression for the evidence, \( P(\mathcal{D}|\mathcal{B}, \beta) \). Writing the expression for the evidence explicitly one obtains

\[
P(\mathcal{D}|\mathcal{B}, \beta) = \pi^{-N/2} |\beta \mathcal{B}|^{1/2} |\beta \psi/2 \pi^{-v/2} \int \prod_{j} dw_{j} e^{-H} \tag{A.1}
\]

where

\[
H = \frac{1}{2} \sum_{jk} r_{j}^{\lambda_{jk}} r_{k} - \sum_{j} \rho_{j} r_{j} + \beta \sum_{\mu} \eta_{\mu}^{2} + \beta \sum_{jk} w_{j}^{0} \mathcal{B}_{jk} w_{k}^{0} \tag{A.2}
\]

while \( r_{j} \equiv w_{j}^{0} - w_{j} \) and

\[
\Lambda_{jk}^{-1} \equiv 2 \beta \left( \frac{1}{N} \sum_{\mu} x_{j}^{\mu} x_{k}^{\mu} + \mathcal{B}_{jk} \right)
\]

\[
\rho_{j} \equiv 2 \beta \left( \frac{1}{\sqrt{N}} \sum_{\mu} x_{j}^{\mu} \eta_{\mu} + (\mathcal{B} w_{j}^{0})_{j} \right)
\]

Performing the gaussian integrals over rule space we obtain

\[
\ln P(\mathcal{D}|\mathcal{B}, \beta) = \frac{1}{2} \ln |\beta \mathcal{B}| + \frac{\rho}{2} \ln \left( \frac{\beta}{\pi} \right) + \frac{1}{2} \ln |\Lambda| + \frac{N}{2} \ln 2
\]

\[
- \beta \sum_{\mu} \eta_{\mu}^{2} - \beta \sum_{jk} w_{j}^{0} \mathcal{B}_{jk} w_{k}^{0} + \frac{1}{2} \sum_{jk} \rho_{j} \Lambda_{jk} \rho_{k} \tag{A.3}
\]

The generalisation error \( \epsilon_{g} \), defined by Eq. 2.7, may be rewritten in terms of the
Figure 1: The generalization error $\epsilon_g$ as a function of the number of patterns $\alpha$ for several values of the decay matrix parameter $\lambda_1$ where $\lambda_2$ is fixed to its optimal value and where $\kappa = 1/2$, $\nu_1 = 0.1$ and $\nu_2 = 1$. The minimal value for $\epsilon_g$ is achieved for the calculated optimal values and typical behaviour for undervalued and overevaluated priors can be seen for the extreme cases.
Figure 2: The generalization error $\epsilon_3$ as a function of the number of patterns $\alpha$ for various choices of $\kappa$, where $\lambda_1 = 5$ and $\lambda_2 = 0.1$, $\nu_1 = 0.1$ and $\nu_2 = 1$. Note the characteristic behaviour of $\epsilon_3$ for undervalued parameters, a hump, which becomes dominant as $\kappa$ decreases. Better performance is achieved in this case when $\kappa$ is larger.
Figure 3: The generalization error $\epsilon_g$ as a function of the number of patterns $\alpha$ for various choices of $\lambda_o$ where $\lambda_2$ is set to its optimal value, $M = 5$, $\mu_1 = 3$ and $\mu_2 = 10$. The minimal value for $\epsilon_g$ is achieved for the calculated optimal value of $\lambda_o$. For low values of $\lambda_o$ the prior over-restricts the data, resulting in a divergence (and a hump for sub-critical values).
Figure 4: The dependence of $\epsilon_3$ on the number of patterns $\alpha$ for several values of the group size $M$ where $\mu_1 = 1$, $\mu_2 = 10$ and the decay parameter $\lambda_d$ and $\lambda_o$ are fixed to their optimal values. The minimal value for $\epsilon_3$ is achieved for the optimal value $M = 11$. 
For deriving \( \epsilon \) one should therefore calculate \(< r_j >_w\). This can be easily done by using the relation

\[
< r_j >_w = \frac{\partial P(D|B, \beta)}{\partial \rho_j} = \sum_k \Lambda_{jk} \rho_k
\]  

(A.5)

Averaging over the test vectors one obtains an expression of the form

\[
< \left[ < \sum_{j=1}^{N} r_j x_j >_w \right]^2 >_x = \sum_{i,j,k,l} \rho_i \Lambda_{ij} \varepsilon_{jk} \Lambda_{kl} \rho_l .
\]  

(A.6)

Averaging over the noise one obtains Eq. 2.8.

Calculating the quenched average \(< (A + L)^{-1} \rangle\) is slightly more complicated and will not be described here in detail. The method presented by Hertz et al (1989) makes use of the expansion

\[
< (A + L)^{-1} > = L^{-1} \left[ I - < AL^{-1} > + < (AL^{-1})^2 > + \cdots \right] ,
\]  

(A.7)

where \( I \) is the identity matrix, to carry out the averages directly and to omit terms of lower orders. Summing up the remaining contributions one obtains Eq. 2.10.
B Explicit solutions for $G$ for the two component penalty matrix

In the case of two diagonal terms, $G$ is the solution of Eq.(2.11). Rewriting Eq.(2.11) explicitly one obtains:

$$G = \frac{\kappa}{\lambda_1 + \frac{\alpha}{1+G}} + \frac{1 - \kappa}{\lambda_2 + \frac{\alpha}{1+G}} \quad (B.8)$$

where $\lambda_1$ and $\lambda_2$ are the two elements along the diagonal of $\mathcal{B}$ and $\kappa = \frac{k}{N}$ and $1 - \kappa$ are their fractions in the diagonal. Eq.(B.8) can be written as a cubic equation:

$$G^3 + a_2G^2 + a_1G + a_0 = 0 \quad (B.9)$$

where

$$a_0 = -\frac{\alpha}{\lambda_1\lambda_2} - \frac{\kappa}{\lambda_1} - \frac{1 - \kappa}{\lambda_2}$$

$$a_1 = 1 + \frac{\alpha - 2\kappa}{\lambda_1} + \frac{\alpha - 2(1 - \kappa)}{\lambda_2} + \frac{\alpha(\alpha - 1)}{\lambda_1\lambda_2}$$

$$a_2 = 2 + \frac{\alpha - \kappa}{\lambda_1} + \frac{\alpha - (1 - \kappa)}{\lambda_2}$$

It has a single positive real root since $a_0$ is always negative, $a_2 \geq 0$ for $\alpha > 1$ and $a_1 < 0$, $a_2 < 0$ for $\alpha < 1$. The real solution is of the form $G = 2\sqrt{p_1}\cos(\theta/3)$ --
$$\alpha_2/3, \text{ where}$$

\[
p_1 = \frac{a_2^2}{9} - \frac{a_1}{3}
\]

\[
p_2 = \frac{a_1 a_2}{6} - \frac{a_0}{2} - \frac{a_2^3}{27}
\]

\[
\theta = \cos^{-1} \left( \frac{p_2}{p_1^{3/2}} \right)
\]