Proceedings of the International Joint Conference on Neural Networks (IJCNN 2002)

On model selection and the disability of neural networks to decompose tasks

Marc Toussaint November 8, 2005

Institut für Neuroinformatik, Ruhr-Universität Bochum, ND 04, 44780 Bochum—Germany
mt@neuroinformatik.ruhr-uni-bochum.de

Abstract

A neural network with fixed topology can be regarded as a parametrization of functions, which decides on the correlations between functional variations when parameters are adapted. We propose an analysis, based on a differential geometry point of view, that allows to calculate these correlations. In practise, this describes how one response is unlearned while another is trained. Concerning conventional feed-forward neural networks we find that they generically introduce strong correlations, are predisposed to forgetting, and inappropriate for task decomposition. Perspectives to solve these problems are discussed.

I Introduction

Following Kerns et al. (1995), the problem of model selection may be defined as follows: Given a finite set of data points, find a function (or conditional probability distribution, also called hypothesis) such that the expected generalization error is minimized. Typically, the search space $\mathcal{F}$ (the space of functions or conditional probability distributions) is assumed to be organized as a nested sequence of subspaces $\mathcal{F}_1 \subseteq .. \subseteq \mathcal{F}_d \subseteq .. \subseteq \mathcal{F}$ of increasing complexity. For instance, the index d may denote the number of parameters or the Vapnik-Chervonenkis dimension (Vapnik 1995). Finding the function with minimal generalization error then amounts to finding the appropriate sub-search-space before applying ordinary optimization schemes. Many approaches introduce a penalty term related to complexity which has to be minimized together with the training error. Penalty terms are, for example, the number of parameters of the model, the number of *effective* model parameters, the Vapnik-Chervonenkis dimension, or the description length (Akaike 1974; Amari 1993; Moody 1991; Rissanen 1978; Vapnik 1995). An alternative based on geometric arguments is presented by Schuurmans (1997).

The emphasis of our investigations is different to these classical approaches. The choice of a specific model (e.g., a neural network) to represent a function has *two* implications: it defines the space $\mathcal{F}_d$ of representable functions, but it also defines a *parametrization* of this space, where parametrization is not meant in the sense of 'finding parameters' but in the sense of introducing coordinates on that space, i.e., introducing a mapping $\Phi : \mathbb{R}^m \to \mathcal{F}_d$ from some coordinate space $\mathbb{R}^m$ onto the sub-search-space. To omit confusion, we use the term *model class* for the sub-search-space $\mathcal{F}_d$, and *model parametrization* for the parametrization Φ of this sub-search-space. For example, an artificial neural network with m free parameters, fixed topology, and fixed activation functions defines a model class (the subspace of functions it can realize—which, if the topology is appropriate, includes an approximation of any function (Hornik, Stinchcombe, & White 1989)) but it also defines a model parametrization (the mapping from its parameters to the corresponding function).

Our emphasis is on the implications of a specific model parametrization instead of the choice of a certain model class. It is important to have a closer look at this parametrization in order to allow for an analytical description of the adaptation dynamics, rather than just analyzing the complexity of a model class. In particular, the precise relation between variations of parameters and functional variations of the system is of fundamental interest because it decides, e.g., on the way of "extrapolation", or on how the system forgets previously learned data. This relation can be derived from the model parametrization and our goal is to extract such features analytically. We focus on forgetting as a specific character of adaptation dynamics and develop an analysis of the model parametrization that allows to approximate the rate of forgetting. This analysis is based on a differential geometry point of view and is related to a large pool of research, including the discussions of *cross-talk* (Jacobs, Jordan, & Barto 1990) and *catastrophic forgetting* (French 1999), the information geometry point of view on parameter adaptation (Amari 2000), and perfectly analogous ideas in the context of evolutionary adaptation (Toussaint 2001). Section III includes a discussion of these relations.

We apply our method of analyzing the model parametrization on the class of standard feed-forward neural networks (FFNNs). We find that the variety of

FFNNs with arbitrary topology is actually not a great variety with respect to certain characters of the model parametrization. In particular, FFNNs gnerically introduce strong correlations between functional variations and thereby are predisposed to forget previously learned data. Hence, using FFNNs as a function model means a limitation—not with respect to representable functions but with respect to learning characteristics. A simple example compares a standard FFNN with a network that includes competitive interactions. The results validate our analytical predictions and illustrate their implications. We conclude that a generalization of the class of FFNNs is necessary and that the introduction of competitive interactions between neurons is a promising approach to solve these problems.

Section II will introduce to the formalism our investigations are based on and, in section III, we describe the analysis of the model parametrization. Section IV presents the examples and in section 5 we give an outlook concerning the evolutionary perspective on model selection and discuss the relevance of the limitedness of FFNN models. The conclusion follows up.

II Definitions

II.1 The functional point of view

Let $\mathcal{F}$ be the search space. Here, $\mathcal{F}$ shall be the space of all functions mapping from a finite space X to $Y \subseteq \mathbb{R}^n$. However, all results can be transferred to the search space of conditional probabilities, as we discuss below.

The space of functions $f : X \to Y$ can be written as Y^X, which is isomorphic to $\mathbb{R}^{n \cdot |X|}$. Thus, let a function $f \in Y^X$ be represented by $n \cdot |X|$ components $f^a \in \mathbb{R}$, where the index a refers to a specific point in X *and* a Y-dimension. (The components f^a may be regarded as entries of a lookup-table representation of f.) On this representation, we describe an online adaptation step as a probabilistic transition to a new function as follows: Assume that adaptation is initiated by the observation of a target value t^a for a functional component f^a. A transition occurs as a variation $\delta f \in \mathbb{R}^{n|X|}$ with probability $p(\delta f \mid f^a, t^a)$. The interesting point is that functional components of which no target value has been observed may vary as well. Let a be a random variable and consider the density $p(\delta f) = p(\delta f \mid f^a, t^a)\, p(a)$. We will refer to the respective covariance between two variation components as the *functional covariance matrix*

$$C^{bc} := \mathrm{cov}_{p(\delta f)}(\delta f^b, \delta f^c) \; . \tag{1}$$

This matrix is a first order description of how the adaptation of the observed functional component results in a *coadaptation* of a functional component which has not been observed. For example, assuming a linear dependence between δf^a and δf^b, we have $\delta f^b \doteq \langle \delta f^b \rangle + \frac{C^{ab}}{\sigma^2} \left(\delta f^a - \langle \delta f^a \rangle \right)$, where σ^2 is the variance of δf^a. Whether this coadaptation is desirable or not depends on the problem. Coadaptation is also an explicit description of the "way of generalization"[1]: unobserved functional components (i.e., the functional response on stimuli that have not been observed) are coadapted depending on the adaptation of observed functional components. In general, one would like to choose from a variety of different coadaptation schemes, i.e., one would like to select a model from a variety of models with different kinds of coadaptation. We will find that this refers to the selection of a model parametrization.

When the set of functional components can be separated in two disjoint subsets such that C^{ab} vanishes for two components f^a and f^b of different subsets, then we speak of *adaptation decomposition*. During online learning, adaptation decomposition means that the development of two such components during successive adaptation is not correlated. In terms of homogeneous Markov processes, successive adaptation is described by the transition probability $p(\delta f \mid f^a, t^a)$ (assuming that the draw of a from $p(a)$ is independent at each time), and adaptation is decomposed if $p(\delta f^a, \delta f^b) = p(\delta f^a)\, p(\delta f^b)$.

II.2 The parameter point of view

We now address the *modeling* of functions. Let Φ be a m-dimensional, differentiable parametrization of a subset $\Phi(W)$ of functions:

$$\Phi : W \to \mathcal{F} \, , \quad W \subseteq \mathbb{R}^m \, , \tag{2}$$

$$\Phi(W) := \bigcup_{w \in W} \{\Phi(w)\} \quad \subseteq \mathcal{F} \, . \tag{3}$$

We call Φ the *model parametrization* and $\Phi(W)$ the *model class*. In terms of differential geometry, Φ is the inverse of a coordinate map (or chart, or atlas) for $\Phi(W)$. Since this map is differentiable, it induces a metric on $\Phi(W)$ if one on W is given and vice versa.

[1] By "way of generalization" we do not refer to the generalization error but to the way of extrapolation from observed data to unobserved.

We define the *functional metric* $g^{ab}(w)$ on $\Phi(W)$ as the lift of the Euclidean metric on W,

$$g^{ab}(w) := \sum_i \frac{d\Phi(w)^a}{dw^i} \frac{d\Phi(w)^b}{dw^i} \; ; \tag{4}$$

and we define the *parameter metric* $g_{ij}(w)$ on W (actually on the *dual* tangent spaces of W) as the pullback of the Euclidean metric on $\Phi(W)$,

$$g_{ij}(w) := \sum_a \frac{d\Phi(w)^a}{dw^i} \frac{d\Phi(w)^a}{dw^j} \; . \tag{5}$$

As usual in differential geometry, the metrics depend on the locality given by w. These metrics describe the relation between parameter variations and functional variations as we explore in more detail in the next section.

III Analysis of the model parametrization

In the previous section we defined the correlation matrix C^{ab} on the functional level. Now we analyze what the choice of a model parametrization Φ implies on this functional level. Given Φ and parameters w, we write $f^a = \Phi(w)^a$. Assume that a target t^a was observed and adaptation of the parameters takes place by a gradient descent,

$$\delta w^i = 2\alpha \, \frac{df^a}{dw^i} \, (t^a - f^a) \; , \tag{6}$$

which corresponds to the gradient of the squared error multiplied by an adaptation rate α. In first order approximation, this induces a functional variation

$$\delta f^b = 2\alpha \, \sum_i \frac{df^b}{dw^i} \, \delta w^i = 2\alpha \, g^{ab} \, (t^a - f^a) \; , \tag{7}$$

using definition (4). Thus, the functional metric g^{ab} describes the variation of a functional component f^b when t^a is observed. This gives a first order description of coadaptation and of how the model generalizes the experience of a target value t^a in order to adapt also functional components f^b. In this approximation the functional covariance reads

$$C^{bc} = 4\alpha^2 \, \sum_a p(a) \, g^{ba} \, g^{ca} \, (t^a - f^a)^2 - \langle \delta f^a \rangle \langle \delta f^b \rangle \; . \tag{8}$$

To discuss this expression, let us assume that the second term vanishes, $\langle \delta f^a \rangle \langle \delta f^b \rangle = 0$. Concerning the first term, the product $g^{ba} \, g^{ca}$ vanishes for all a if and only if the functional metric is a block matrix and b and c refer to different blocks:

$$g^{ab} = \begin{pmatrix} A \in \mathbb{R}^{\mu \times \mu} & 0 \\ 0 & B \in \mathbb{R}^{\nu \times \nu} \end{pmatrix} \; , \quad b \leq \mu \; , \quad c > \mu \; ,$$

where A and B are arbitrary symmetric matrices and $\mu + \nu = n \cdot |X|$. Thus, adaptation is decomposed into two subsets of functional components exactly if the functional metric is a block matrix and the functional component subsets correspond to these blocks.[2]

III.1 Reference to related research

Cross-talk. The inspiring work by Jacobs et al. (1990) discusses the implication of the choice of a multi-expert model on the learning speed and generalization behavior. They formulate the idea of spatial and temporal crosstalk, which denotes the statistical dependence between the states of two different neurons or between the states of a neuron at two different times. In our formalism, this crosstalk is captured by the functional covariance—spatial for two indices a and b belonging to the same input $x \in X$, and temporal for two indices of different input. They argue that such a crosstalk may be undesirable and is avoided by explicitly separating neurons in disjoint experts. As we will see below, selecting a multi-expert model is a very intuitive way to explicitly declare an independence of functional components and realize decomposed adaptation. In fact, the separation into experts corresponds to a block matrix type functional metric. (If the gating is also adaptive, the functional metric is actually not a completely clean block matrix.)

In the context of artificial neural networks, the term *catastrophic forgetting* has been used to describe negative effects of coadaptation. See (French 1999) for a review.

[2] Note the relation to group theory: A group representation is said to be reducible if all group generators can be represented as a block matrix (such that all of them fit in the same block template). On this basis, physics defines the notion of an elementary particle as corresponding to an irreducible representation, whereas physical systems that correspond to a reducible representation (a block matrix) are considered as *composed* of particles. A system of which the adaptation dynamics (instead of physical interactions) can be decomposed in the sense of a block matrix can analogously be thought of as composed of subsystems.

More formally, the observation of a target t^a can be identified with an element of a group that applies on the functional components. Adaptation dynamics is now interpreted as successive application of group elements. The group representation (i.e., the way the group elements apply on the functional components) is determined by the model parametrization. If adaptation is decomposed, this representation is reducible.

Information geometry. The methods applied in this paper are related to information geometry. Let $Y = S_\nu = [0,1]^{2^\nu - 1}$ be the $2^\nu - 1$ dimensional manifold of probability distributions over $\{0,1\}^\nu$ as defined by Arami (2000). Then, the search space $\mathcal{F}$ of mappings $X \to Y$ is the space of all conditional probabilities $p(y|x)$, $x \in X, y \in Y$. Usually, one assumes the Fisher metric on $\mathcal{F}$, not the Euclidean. Thus, we would have to change the definition (5) of the parameter metric into

$$g_{ij}(w) = E\left[\frac{\partial \log p(x,y;w)}{\partial w^i}\,\frac{\partial \log p(x,y;w)}{\partial w^j}\right] , \tag{9}$$

where $E[.]$ denotes the expectation and $p(x,y;w) = p(y|x;w)\,p(x)$, $p(y|x;w) = \Phi(w) \in \mathcal{F}$. Arami (1998) uses this metric to define the natural gradient descent on the parameter space (which actually is the covariant derivative instead of the contravariant). The use of the natural gradient can also be motivated by a spatio-temporal decorrelation (Choi, Amari, & Cichocki 2000).

Evolutionary computation. It seems that in the field of evolutionary computation the discussion of the covariance structure in the search space is much more elaborated than in the field of neural computation (see Toussaint 2001). Roughly speaking, the goal of evolutionary computation is to maximize the probability of good mutations during evolutionary search. Eventually, fitness requires some phenotypic traits to be mutated in correlation. Such correlations (coadaptation) may be modeled explicitly in the search density of evolutionary algorithms (Baluja & Davies 1997; Hansen & Ostermeier 2001; Mühlenbein, Mahnig, & Rodriguez 1999; Pelikan, Goldberg, & Lobo 1999). Alternatively, they may be induced implicitly by the choice of a good parametrization of phenotypic traits—by a genotype-phenotype mapping, which is in perfect analogy to the model parametrization Φ. Many research efforts focus on the choice or the understanding of the genotype-phenotype mapping (Stephens & Waelbroeck 1999; Toussaint 2001; Wagner & Altenberg 1996). In this view, functional components f^a may be compared to phenotypic traits, whereas parameters relate to the genotype.

IV Example

Our test of the learning behavior is very simple: a regression of only two patterns in $\{0,1\}^3$ has to be learned by mapping the first pattern on $+1$ and the second on -1. However, we impose that these patterns have to be learned *online* where they alternate only after they have been exposed for 100 times in succession.[3] We test two systems on this task: a standard feed-forward neural network as described in detail in table 1, and a system that involves a softmax layer as described in table 2. The parameters of both systems are initialized randomly by the normal distribution $\mathcal{N}(0, 0.1)$ around zero with standard deviation 0.1. The two patterns were chosen as 110 and 010. Learning is realized by a slow gradient descent with adaptation rate $2 \cdot 10^{-3}$ and momentum 0.5. The metric components are calculated from the gradients.

The feed-forward neural network we investigate here is 3-4-1-layered; layers are completely connected; the output neurons are linear, the hidden ones implement the sigmoid $\frac{1}{1+\exp(-10\,x)}$; only the hidden neurons have bias terms.

Table 1: The Standard model

The softmax model is the same as the standard model with the exception that the four neurons in the hidden layer *compete* for activation: their output activations y_i are given by

$$y_i = \frac{e^{30\,x_i}}{X} , \quad x_i = \sum_{j \in \text{ input}} w_{ij} y_j + w_i ,$$
$$X = \sum_{i \in \text{ hidden}} e^{30\,x_i} . \tag{10}$$

Here, w_{ij} and w_i denote weight and bias parameters. The exponent factor 30 may be interpreted as rather low temperature, i.e., high competition. The calculation of the gradient is a little more involved than ordinary back-propagation but straightforward and of same computational cost (see (Toussaint 2002)).

Table 2: The Softmax model

Please see Figures 1 and 2 for the results. For the standard neural model we observe some forgetting of the untrained pattern during the training of the other. For the softmax model, the error of the untrained pattern hardly increases. The rate of for-

[3] This task is not meant as a performance test but as an experimental setup to test our analytical methods. However, similar effects of learning and unlearning occur in online learning when a specific response is unlearned during the course of training other responses for several time steps. In real world simulations it is also plausible that stimuli remain unchanged for many time steps.

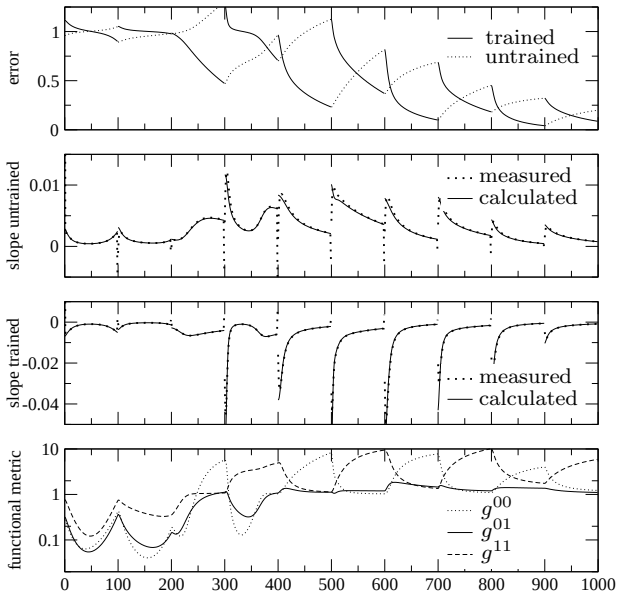

Figure 1: *Test of the standard model.*
For all four graphs the abscissa denotes the time step.
Top: The learning curves (errors) with respect to both patterns are displayed. Only one of the patterns is trained—alternating every 100 time steps. The error of the untrained patterns increases.
Second: The slope (change of error per time step) of the untrained learning curve is displayed. The dotted line refer to the measured slope of the upper curve, the normal line is calculated according to equation (7).
Third: The slope (measured and calculated) of the trained learning curve.
Bottom: The three components of the functional metric g^{00}, g^{01}, g^{11} are displayed in logarithmic scale. In particular the cross-component g^{01} is clearly non-vanishing.

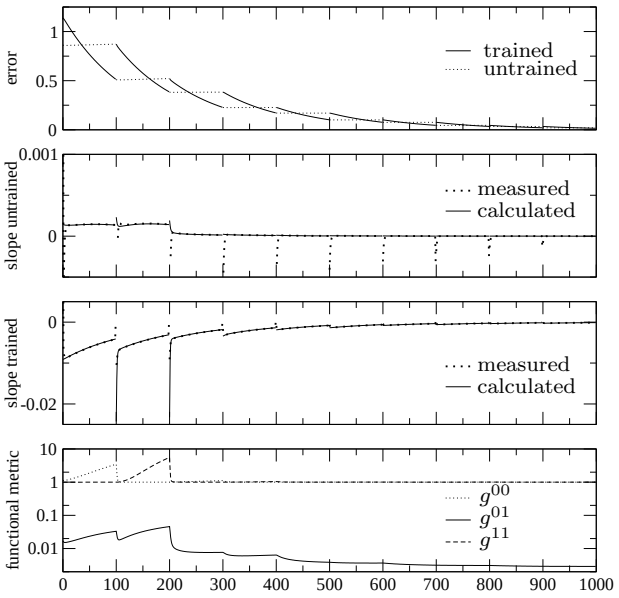

Figure 2: *Test of the softmax model.*
Top: The learning curves (errors) with respect to both patterns are displayed. The untrained patterns is scarcely forgotten.
Second: The slope (measured and calculated) of the untrained learning curve nearly vanishes.
Third: The slope (measured and calculated) of the trained learning curve.
Bottom: The three components of the functional metric g^{00}, g^{01}, g^{11} (in logarithmic scale). The cross-component g^{01} is small, it decreases significantly at time step 200.

getting, given by the slope of the error curve, is well described by equation (7) and demonstrated by the graphs in the middle. The bottom graphs display the functional metric components and generally exhibit that the cross-component g^{01}, which is responsible for coadaptation and forgetting, is quite large for the standard model compared to the softmax model. Further, the softmax model seems to learn the adaptation decomposition, as defined in section II, after the 200th time step. All these results reveal that the standard model is not well-suited to solve the simple task given and that the analysis of the model's functional metric provides a formal way of understanding this phenomenon. Remarkably also, the components g^{00} and g^{11} become significantly greater than 1 during the training phase of the respective functional component. By equation (7), this means that the "effective" adaptation rate is larger than $2 \cdot 10^{-3}$.

One might object that the results given above rely on the random initialization and on the specific task we chose. To analyze both types of models in a more general way we perform another test. We investigate the distribution of the functional metric components when parameters are normally distributed by $\mathcal{N}(0, 0.1)$. Figure 3 shows the distributions for both models. Clearly, the standard model exhibits a Gauss-like distribution of the cross-component g^{01} with mean around 1.5; a vanishing cross-component g^{01} is not very likely. On the other hand, the softmax model exhibits two strong peaks at $g^{01} = 0$ and $g^{01} = 1$, such that the probability for $g^{01} < 0.1$ is larger than 10%. These distributions are generic properties of the two models.

V Toward evolutionary model selection

Finally, the question of how to select an appropriate model has not yet been addressed. As discussed in the introduction, classical approaches to model selection commonly introduce a penalty term in order to reduce the model's complexity. Following this tradition we could introduce a penalty term that reduces

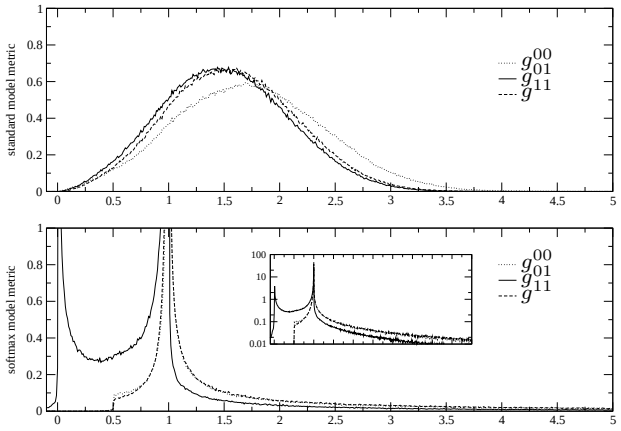

Figure 3: *Distribution of metric components.*
The distribution was calculated as a histogram of 1 million samples by using bins of size $\frac{1}{100}$. The ordinate is scaled in "percent of samples that fell into the bin".
Top: The standard model. The probability of vanishing cross-component g^{01} is vary small.
Bottom: The softmax model. The inset graph is in logarithmic scale. The probability of vanishing cross-component g^{01} is fairly high.

forgetting. Consider

$$\sum_{ab}(g^{ab})^2 - \sum_{a}(g^{aa})^2 \ . \tag{11}$$

This is a measure of the cross-components in the functional metric. Unfortunately, we cannot present any experiments with this model selection criterion here. This approach is postponed to future research.

The original motivation for this work, though, was not to develop a new model selection criterion as given by the above penalty term. Instead we believe that the evolution of neural networks, as it recently became an elaborated branch of research (see (Yao 1999) for a review), is actually a promising method of model selection. However, most of these approaches focus on standard neural models, i.e., the evolutionary search space is the space of ordinary feed-forward neural networks (FFNNs) with arbitrary topology. The belief is that the variety of topologies offers a variety of functionally different models. The present paper is a critique of this belief because it supports that the functional metric inherent of FFNNs comprises significantly non-vanishing cross-components. This implies that the variety of FFNNs with arbitrary topology is actually not a great variety with respect to the functional metric. E.g., it hardly includes models with vanishing cross-components and low rate of forgetting. In conclusion, the search space has to be generalized to contain also models with arbitrary functional metric in order to allow for the selection of more optimal models. The presented softmax model involving competitive interactions between neurons is a step in this direction, but much motivation is left for future research toward the generalization of the model search space and evolutionary methods to select good models from this great variety. The model presented in (Toussaint 2002) is one approach.

VI Conclusion

We developed a new analytical approach to characterize a function model and describe its learning properties. We focussed on functional correlations in the adaptation process and derived the relation to the functional metric of the model parametrization. The analysis can in principal be applied on any kind of differentiable model (also probabilistic, when formulated in terms of information geometry). Our empirical studies illustrate the approach and demonstrate that conventional neural network models are rather limited with respect to their adaptation behavior: a task separation, i.e., decorrelated adaptation to decorrelated data, is hardly possible. In contrast, a model involving competitive interactions is more predisposed for task decomposition. Thus, as we pointed out in the previous section, the evolutionary approach to model selection should generalize the search space to include not only standard feed-forward neural networks, but also models with arbitrary functional metrics, e.g., by allowing for competitive interactions.

Acknowledgment

The author acknowledges support by the German Research Foundation DFG under grant *SoleSys* SE 251/41-1.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC–19**, 716–723. For a reprint see E. Parzen et al. (Eds.), *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, 1998.

Amari, S. (1993). Mathematical methods of neurocomputing. In O. Barndorff-Nielsen, J. Jensen, & W. Kendall (Eds.), *Networks and Chaos—Statistical and Probabilistic Aspects*, pp. 1–39. Chapman & Hall, London.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation* **10**, 251–276.

Amari, S. (2000). Information geometry on hierachical decomposition of stochastic interactions. Submitted to IEEE Transactions on Information Theory.

Baluja, S. & S. Davies (1997). Using optimal dependency-trees for combinatorial optimization: Learning the structure of the search space. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML-97)*, pp. 30–38.

Choi, S., S. Amari, & A. Cichocki (2000). Natural gradient learning for spatio-temporal decorrelation: recurrent network. *IEICE Transactions Fundamentals* **E83**, 2715–2722.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Scienences* **3**, 128–135.

Hansen, N. & A. Ostermeier (2001). Completely derandomized self-adaption in evolutionary strategies. *Evolutionary Computation* **9**, 159–195.

Hornik, K., M. Stinchcombe, & H. White (1989). Multilayer feedforward networks are universal approximators. *Neural Networks* **2**, 359–366.

Jacobs, R. A., M. I. Jordan, & A. G. Barto (1990). Task decomposistion through competition in a modular connectionist architecture: The what and where vision tasks. Technical Report COINS-90-27, Department of Computer and Information Science, University of Massachusetts Amherst.

Kearns, M., Y. Mansour, A. Y. Ng, & D. Ron (1995). An experimental and theoretical comparison of model selection methods. In *Proceedings of the Workshop on Computational Learning Theory (COLT-95), Morgan Kaufmann Publishers*.

Moody, J. (1991). The effective number of parameters: an analysis of generalization and regularization in nonlinear systems. In *Advances in Neural Information Processing Systems*, Volume 4, pp. 847–854.

Mühlenbein, H., T. Mahnig, & A. O. Rodriguez (1999). Schemata, distributions and graphical models in evolutionary optimization. *Journal of Heuristics* **5**, 215–247.

Pelikan, M., D. E. Goldberg, & F. Lobo (1999). A survey of optimization by building and using probabilistic models. Technical Report IlliGAL-99018, Illinois Genetic Algorithms Laboratory.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica* **14**, 465–471.

Schuurmans, D. (1997). A new metric-based approach to model selection. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, pp. 552–558.

Stephens, C. & H. Waelbroeck (1999). Codon bias and mutability in HIV sequences. *Molecular Evolution* **48**, 390–397.

Toussaint, M. (2001). Neutrality and self-adaptable exploration in evolutionary search. Technical Report IRINI-2001-05, Institut für Neuroinformatik, Ruhr-Universität Bochum.

Toussaint, M. (2002). A neural model for multi-expert architectures. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2002)*.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer, New York.

Wagner, G. P. & L. Altenberg (1996). Complex adaptations and the evolution of evolvability. *Evolution* **50**, 967–976.

Yao, X. (1999). Evolving artificial neural networks. In *Proceedings of IEEE*, Volume 87, pp. 1423–1447.