Data Compression and Prediction in Neural Networks

Ronny Meir*
Department of Electrical Engineering
Technion
Haifa 32000, Israel
rmeir@ee.technion.ac.il

Jose F. Fontanari†
IFQSC - DFCM
Universidade de São Paulo
13560 São Carlos SP, Brazil
fontanari@uspfsc.ifqsc.usp.ansp.br

Abstract

We study the relationship between data compression and prediction in single-layer neural networks of limited complexity. Quantifying the intuitive notion of Occam's razor using Rissanen's minimum complexity framework, we investigate the model-selection criterion advocated by this principle. While we find that the criterion works well for large sample sizes (as it must for consistency), the behavior for finite sample sizes is rather complex, depending intricately on the relationship between the complexity of the hypothesis space and the target space. We also show that the limited networks studied perform efficient data compression, even in the error full regime.

1 Introduction

Many problems in signal processing and machine learning can be described as making predictions about the future based on the past. In particular, in pattern recognition we are usually given a stream of input/output pairs $D_m = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_m, y_m)\}$ where $\vec{x}_m$ are N-dimensional input vectors and y_m are class labels. We are then usually requested to do one of three things. In the *prediction* problem we are required to predict the value of a label y_{m+1} of a new input example $\vec{x}_{m+1}$. In the *data compression* problem one wishes to find a means to compress the data D_m, so that fewer bits are needed to store it. If the data is not random (i.e. is governed by a rule) we expect to be able to compress it. Third, in the *estimation* problem we wish to construct a model which explains the data. Clearly such a model makes predictions regarding the future, as well as encoding the data. Finally, there is the meta-problem of *model selection*, i.e. given several models to explain the data how do we pick the 'best' model. This problem has a long history in

*Research supported in part by the Ollendorf center of the electrical engineering department at the Technion and by the Lady Davis foundation.
†Research supported in part by the CNPq.

the philosophical literature, but has only been quantified in mathematical terms during the past thirty years. In particular, the approach we take is based on the ideas of Rissanen [Ris86], whose work in turn builds on the results of Shannon [Sha48] and Kolmogorov [Kol65].

In this paper we address the problem of model selection, within a very simple class of models, namely single-layer perceptrons with pre-defined constraints on the weights. While limited in scope, this class turns out to be very productive in affording us insight about the four problems described in the first section, as well as being exactly solvable. Moreover, the model studied allows us to extract the typical number of bits needed to store a set of input/output relationships in a single-layer perceptron, a quantity with an interest of its own.

2 Data Compression and Prediction

Before plunging into the details of our specific model we briefly review some issues related to data compression in the context of learning and pattern recognition. We assume as before a data set consisting of a stream of m input/output pairs

$$D_m = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_m, y_m)\}. \tag{1}$$

We also find it convenient to use a shorthand notation for the input set, $x^m = \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m\}$ and $y^m = \{y_1, \ldots, y_m\}$. Although not needed for the general discussion we assume that the outputs y_i are binary ± 1 variables. We assume throughout this section that the input set x^m is fixed, while the output sequence y^m is picked according to some unknown rule f, i.e. $y_i = f(\vec{x}_i)$. Since the stream of outputs y^m is not random it can therefore be represented more efficiently than by simply presenting the actual values y^m, which require m bits. In order to construct an efficient representation of the data one usually constructs a *model* h (for hypothesis) and tries to fit the data to the model. We assume that the model h is chosen from some space $\mathcal{H}$ of possible hypotheses. Each hypothesis usually consists of a parameterized set of functions. For example, in the context of neural networks, h can represent all neural networks with a single hidden layer and no more than 77 connections. The parameters in this case are just the weights and thresholds of the neurons. One then usually goes through an *estimation* stage where the 'best' parameter for a given hypothesis is chosen. Later, the predictions of different models are compared and the 'optimal' model is chosen. Now, as has been pointed out by Rissannen [Ris86] this process can be viewed as one of encoding the data according to a two-level scheme. In the first stage we encode the description of the parameters themselves while in the second we encode the data given the actual values of the parameters. The sum of these two quantities is exactly the number of bits needed to specify the data using hypothesis h. Rissanen's *minimum description length* criterion then instructs us to pick that model with the lowest overall description length. The appealing feature of this criterion is that it is based solely on the training set, and does not require any evaluations of data outside of the training set. Rissanen has shown that under certain conditions, this two-level process yields *asymptotically* (for large sample size) the best predictor. While nice as a consistency demonstration this result does not tell us about the behavior of the method for *finite* sample sizes. It is the explicit aim of this paper to investigate this criterion in an exactly solvable case. Not surprisingly we find that for finite sample sizes the problem becomes much more complicated and depends sensitively on the complexity level of the learner as well as that of the target rule.

3 Statistical Physics and Data Compression

We consider a neural network with N inputs, parameterized by a weight vector $\vec{w}$. The output of the network is given by a functional relationship $f_w(\vec{x})$. We further define an error-function $E(\vec{w}, D_m)$ measuring the accumulated deviation between the desired and actual outputs taken over the training set D_m (in the calculations presented in this work we take E to represent the number of misclassified examples [GD88]). Within the statistical physics framework, learning takes place through a modification of the probability distribution on weight space due to incoming data [LTS90]. The posterior distribution, which can easily be derived from the maximum entropy principle, gives rise to the Gibbs distribution

$$P(\vec{w}|D_m, h) = \frac{e^{-\beta E(\vec{w}, D_m)}}{Z_m(D_m)}, \tag{2}$$

where the partition function Z_m is given by

$$Z_m(D_m) = \int d\mu(\vec{w}) e^{-\beta E(\vec{w}, D_m)} \quad . \tag{3}$$

We use the generic symbol $d\mu(\vec{w})$ to indicate the measure on weight space, be it discrete or continuous, although we concern ourselves here with discrete measures. To simplify the notation we have dropped the explicit dependence on h. It should be kept in mind however, that all quantities calculated within the statistical mechanics approach refer to a particular model h.

To quantify the notion of prediction we define $P(y_{m+1}|D_m, \vec{x}_{m+1})$, the probability of predicting y_{m+1}, given input $\vec{x}_{m+1}$, *after* having being exposed to m input-output examples D_m. Integrating over the posterior distribution of weights, $P(\vec{w}|D_m)$, we find

$$P(y_{m+1}|D_m, \vec{x}_{m+1}) = \int d\mu(\vec{w}) P(y_{m+1}|\vec{x}_{m+1}, \vec{w}) P(\vec{w}|D_m) = \frac{Z_{m+1}(D_m, \vec{x}_{m+1}, y_{m+1})}{z Z_m(D_m)}, \tag{4}$$

where we have used $P(y|x, \vec{w}) = \exp(-\beta e(y|x, \vec{w}))/z$, with $e(y|\vec{w}, \vec{x})$ representing the single pattern error [LTS90]. The denominator z normalizes the probability so that $\sum_y P(y|x, \vec{w}) = 1$. For the error function used in this work, $e(y|x, w) = \Theta(-y f_w(\vec{x}))$, we have $z = 1 + e^{-\beta}$.

Now, the probability distribution $P(y_{m+1}|D_m, \vec{x}_{m+1})$ allows us to encode the random variable y_{m+1} with the 'ideal' code length $-\log P(y_{m+1}|D_m, \vec{x}_{m+1})$. As described in [Ris86], "The word ideal here means that if the possible values of the next observation are indeed distributed as modeled, then no prefix code[1] exists with a shorter mean length".

To obtain the total code length for the data, given the model h, we sum over the optimal code lengths just described, obtaining the so-called *stochastic complexity* [Ris86]

$$I(D_m|h) = -\sum_{l=0}^{m-1} \log P(y_{l+1}|D_l, \vec{x}_{l+1}) = -\sum_{l=0}^{m-1} \log \frac{Z_{l+1}}{Z_l} = -\log \frac{Z_m}{Z_0}. \tag{5}$$

Note that Z_0, the partition function before any data have arrived, is simply equal to the number of configurations in weight space, i.e. $Z_0 = \int d\mu(\vec{w})$. The importance of the stochastic complexity is that it represents *the length in bits of the shortest description of the data, given the model h.* We also note that the stochastic complexity, I, is very closely related to the free-energy. In particular we have

$$I(D_m|h) = \beta F(D_m|h) + \log Z_0. \tag{6}$$

Up to now all quantities depend on the data D_m. In the remainder of the paper we take the usual thermodynamic limit with $N \to \infty$, $m \to \infty$ and $m/N = \alpha$. Moreover, we focus on the *zero temperature* limit $\beta \to \infty$ where the posterior distribution (2) is nonzero only on the subspace of minimal training-error. Using the basic assumption of the statistical physics of quenched disorder, we assume that all extensive quantities are self-averaging in the thermodynamic limit and in particular that

$$I(x^m, y^m|h) \to \langle I(x^m, y^m|h) \rangle_{x^m}, \tag{7}$$

where we have not yet averaged over the output sequence y^m. Now, to obtain average case results we average this quantity also over the labels y^m, obtaining the average code-length for the data given the model,

$$\mathcal{I}(\alpha) \stackrel{\text{def}}{=} \lim_{N,m\to\infty} \frac{1}{N} \langle I(x^m, y^m|h) \rangle_{D_m} = (\mathcal{S}_0 - \mathcal{S}) + \alpha\beta\epsilon_t \tag{8}$$

where $\mathcal{S}_0 = (1/N) \log \int d\mu(\vec{w})$, $\mathcal{S}$ is the entropy density and $\epsilon_t = \langle E(\vec{w}, D_m)/m \rangle$ is the average training error. This expression has a very simple and intuitive interpretation. The first term represents the number of bits needed to describe the model, given the data. For example, for small α there are many weight configurations consistent with the data. Since we only need one of them, it is clear that we need specify very few bits to pin-point one of them. As α increases, the number of configurations of minimal error decreases, yielding the description of the model more lengthy. The second term in (8) is related to the fact that the hypothesis h may not yield an error free description of the data. Thus, in order to describe the data given the model, we have to encode the errors as well. We can therefore interpret the above expression as composed of two parts, the first describing the complexity of the model, while the second gives the description length of the data given the

[1]In a prefix code no codeword is a prefix of any other codeword.

model. We note that this interpretation is very close in spirit to Rissanen's minimum description length principle, although it differs in detail. In particular, Rissanen calculates the code length needed to describe the data given the model, where the parameters of the model are estimated from the *maximum likelihood* principle. He then adds a term containing the description length of the parameters. In our formulation we use the complete posterior distribution in parameter (weight) space and thus automatically get both the code length of the model and the data given the model. The model selection criterion suggested by Rissanen is to to pick the model with the *minimum stochastic complexity* (MSC). It is our explicit aim to investigate how this criterion relates to the generalization error.

An important constraint which every model selection criterion should obey is that it is asymptotically consistent, i.e. if the true model generating the examples is a member of the hypothesis class $\mathcal{H}$, it should ultimately be correctly identified. From the expression for $\mathcal{I}(\alpha)$ it is obvious that for large α the model with lowest training error is to be chosen, as the energy term dominates the entropy. It is impossible that two different models have zero training error in this limit, since if they give the same result for *every* input they must be identical. Even if the true model is not a member of the hypothesis space, it can be seen that the criterion selects the model closest to the true model. To see this note that, as mentioned above, for large α the selected model is the one with the smallest training error. Since in this limit the training and generalization errors converge to a common limit it is clear that the selected model will have the lowest generalization error, i.e. is the closest hypothesis to the true model.

4 Learning in a Perceptron of Bounded Precision

Up to now our discussion has been rather general. In order to make the above ideas more concrete, we study the problem of learning from examples in a single-layer perceptron whose weights are constrained to take on only a limited set of values. There are several reasons for studying this model. (i) Within the coding length framework, it is very natural to think about finite precision weights, since the amount of information in each weight is easily derivable, without worrying about quantization issues. (ii) The representation ability of finite precision networks (in terms of error rate) is only slightly worse than that of a network composed of real weights [MF92]. (iii) From the statistical physics point of view these networks have the added advantage that the replica symmetric solution is locally stable (if calculated within the micro-canonical ensemble [FM93]), implying that it is probably the correct solution. For networks composed of real weights [GD88], one usually finds that replica symmetry needs to be broken, making the exact solution of the problem much more complicated.

Our work then treats a single-layer *student* perceptron whose weights are constrained to take on values $\{\pm 1, \pm 2, \ldots, \pm L_s\}$. Thus, the maximal number of bits needed to specify each weight is $\log(2L_s)$. The *teacher* supplying the labels y^m is assumed to be a single-layer perceptron as well. We study two models: (i) The (L_s, L_t) model where the student weights are constrained as above, and the teacher weights take values in $\{\pm 1, \pm 2, \ldots, \pm L_t\}$. It is clear then that for $L_s \geq L_t$ the problem is *realizable*, while for $L_s < L_t$ it is *unrealizable*. As we show below the behavior of the two models is very different. (ii) The (L_s, R) model where the learner is constrained as above, while the data pairs are produced from a single-layer perceptron whose weights are real numbers. This problem is clearly unrealizable.

Using standard teachniques, pioneered by E. Gardner [Gar88][GD88], we are able to fully solve the above two models. The main features of the solution to the finite precision problem at zero temperature are the following (for a full discussion of these results see [MF92][MF93]). (i) For all problems there is a value of α below which training with zero error is possible (even in the unrealizable problem). The entropy of the system at this value of α vanishes; thus we term this value α_{ZE}. (ii) In the realizable case $L_s \geq L_t$ there is a sharp transition to perfect generalization at α_{ZE} [SST92][MF92]. (iii) For the unrealizable problem above α_{ZE}, there is a freezing of the system at a nonzero temperature $T_{\mathrm{ZE}}(\alpha)$, i.e. the entropy of the system vanishes for all $T \leq T_{\mathrm{ZE}}(\alpha)$. The training error of the network above α_{ZE} is nonzero. (iv) $N \log \alpha$ bits are sufficient to perfectly implement $N\alpha$ input/output mappings generated from a perceptron with real weights [MF93]. Thus, for sample size α, a rather small number of bits is sufficient to characterize the data.

Before proceeding with a detailed discussion of the results we note that the stochastic complexity $\mathcal{I}$ takes on a particularly appealing form in this problem. Since the average training error is zero below α_{ZE} while the entropy is zero above α_{ZE} we have

$$\mathcal{I}(\alpha, T=0) = \begin{cases} \mathcal{S}_0 - \mathcal{S}(\alpha) & \text{if } \alpha \leq \alpha_{\mathrm{ZE}} \quad (\epsilon_t = 0) \\ \mathcal{S}_0 + \alpha\beta_{\mathrm{ZE}}\epsilon_t & \text{if } \alpha > \alpha_{\mathrm{ZE}} \quad (\epsilon_t > 0) \end{cases} \tag{9}$$

This form for $\mathcal{I}$, the minimum description length, has the following intuitive explanation. Below α_{ZE} the model reproduces the data perfectly. Thus to describe the data it is sufficient to specify only the model. As discussed earlier, the minimal cost needed to describe the model is $\mathcal{S} - \mathcal{S}_0$. Above α_{ZE} the cost needed to describe the data is composed of the cost needed to describe the model $\mathcal{S}_0$, as well as the cost needed to specify the errors, namely $\alpha\beta_{\mathrm{ZE}}\epsilon_t$.

Discrete teacher

In this model, referred to as the (L_s, L_t) model, the student and teacher weights are constrained to values $\{\pm 1, \pm 2, \ldots, \pm L_s\}$ and $\{\pm 1, \pm 2, \ldots, \pm L_t\}$ respectively. The teacher weights are chosen uniformly at random from their allowed range. In figure (1a) we plot the generalization error against the sample size α. As mentioned above, for all realizable models, $L_s \geq L_t$, there is a transition to perfect generalization ($\epsilon_g = 0$) at a finite value of α. This can be seen from the lines marked $5, 7, 9$

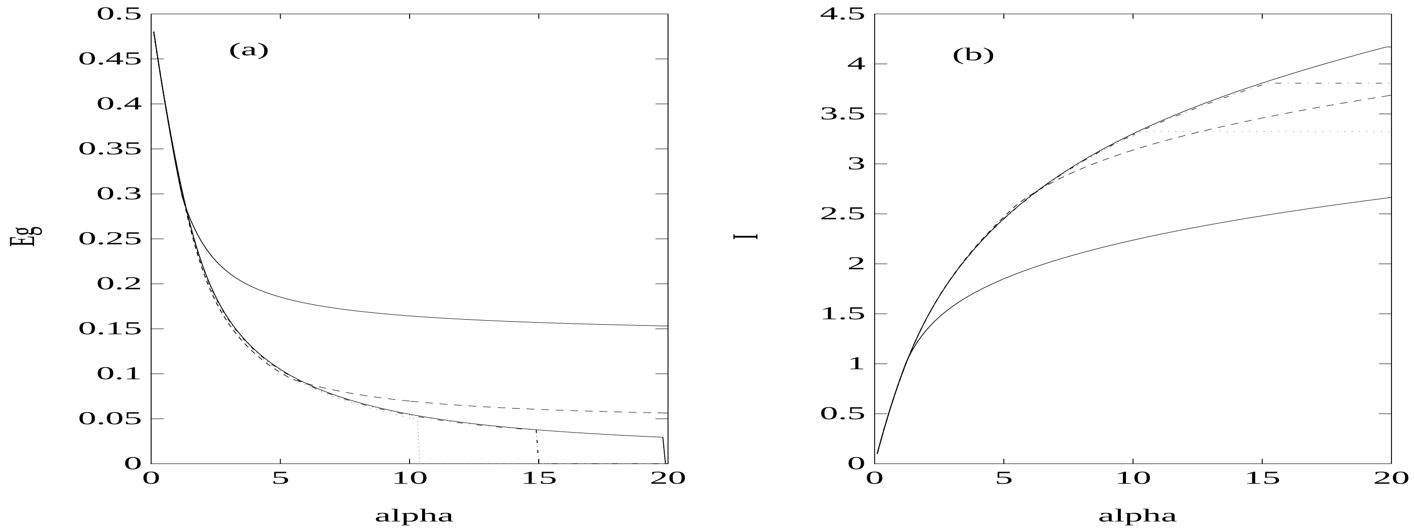

Figure 1: (a) Generalization error for the (L_s, L_t) model with $L_t = 5$ and $L_s = 1$ (solid line), $L_s = 3$ (dashes), $L_s = 5$ (dots), $L_s = 7$ (dot-dash) and $L_s = 9$ (solid line). (b) Same as (a) but for the stochastic complexity.

in the figure. In the unrealizable problem, $L_s < L_t$, the asymptotic generalization error is never zero, but rather approaches a minimal value $\epsilon_{\min}$. We find the following asymptotic behavior

$$\lim_{\alpha\to\infty} \left[\epsilon_g(L_s, L_t, \alpha) - \epsilon_{\min}(L_s, L_t)\right] = \begin{cases} 0 & L_s \geq L_t \\ f(L_s, L_t)\ \alpha^{-1} & L_s < L_t\ . \end{cases} \tag{10}$$

For finite values of α we find that the generalization error displays an interesting behavior. For small α, ϵ_g is lowest for the model with smallest L_s, i.e. $L_s = 1$. As α increases there is a transition to a situation where $L_s = 3$ provides the lowest generalization error . As α increases further the 'true' model $L_s = 5$ takes over in the sense of having lowest generalization error . We note that the over-realizable models, $L_s > L_t$ never possess the lowest generalization error (except for large α where all realizable models yield $\epsilon_g = 0$).

In order to compare this behavior to that predicted by the MSC principle we first present the asymptotic behavior of the stochastic complexity $\mathcal{I}$,

$$\lim_{\alpha\to\infty} \mathcal{I}(L_s, L_t, \alpha) = \begin{cases} \log(2L_s) & L_s \geq L_t \\ \log(2L_s) + g(L_s, L_t)\ \alpha^{0.3} & L_s < L_t\ . \end{cases} \tag{11}$$

It is interesting to note that the minimal description length scales as $\alpha^{0.3}$ which is much smaller than the number of bits needed to specify the positions of the $\alpha N \epsilon_t$ errors, which scales as α. Thus, although the network produces errors on the data, there is an efficient way to encode the data which requires a small number of bits per pattern. We therefore conclude that the network performs real data compression, even in the unrealizable regime.

We plot in figure (1b) the stochastic complexity as a function of sample size α for the different models. As can be seen the behavior of this quantity is not simple. For small α we find that the smallest description length is obtained for the 'true' model $L_s = 5$ (we note that for small α the difference between the various models is $O(\alpha^2)$ and cannot be discerned on the scale of the figure). As α increases it is the simplest model $L_s = 1$ which possesses the shortest description length. Asymptotically of course we find that the $L_s = 5$ model has the shortest description length, although this occurs for α much higher than the largest value in the figure. It is interesting to note that if we plot the generalization error against the stochastic complexity for each value of α (for a particular value of L_s) we find a monotonic relationship. Thus, there seems to be a clear relationship between the two quantities, the exact form of which depends however on the specific details of the problem.

Continuous teacher

We study now the case where the real weights of the teacher perceptron are chosen independently from a gaussian distribution of mean zero and unit variance. Since the number of bits needed to describe a real number is potentially infinite, it is interesting to see what data reduction can be achieved here if the weights are constrained to a finite set of values.

The asymptotic behavior of the generalization error obtained for this model is given by

$$\lim_{\alpha\to\infty} [\epsilon_g(L_s,\alpha) - \epsilon_{\min}(L_s)] = f'(L_s)\ \alpha^{-4/5}. \qquad (12)$$

We note that compared to the unrealizable discrete teacher case ($L_s < L_t$), the relaxation of the generalization error to its minimal value is slower. This is reasonable since the present problem is more difficult to learn by a finite precision network.

The behavior of the generalization error for finite α is depicted in figure (2a) for $L_s = 1, 5, 10$. We find that for every value of α it is always the more complex model (higher L_s) which yields a lower generalization error . This should be contrasted with the previous model where for small α the simpler network yielded lower generalization error .

The asymptotics of the stochastic complexity is given by

$$\lim_{\alpha\to\infty} \mathcal{I}(L_s,\alpha) = \log(2L_s) + g'(L_s)\ \alpha^{0.4}. \qquad (13)$$

Again we note that the number of bits needed to specify the data grows much more slowly than α, the number of bits needed assuming independence. From the actual values calculated for $g'(L_s)$ we find for large α, $\mathcal{I}(1) > \mathcal{I}(5) > \mathcal{I}(10)$ as we would expect, since the $L_s = 10$ model is closest to the true model. As mentioned at the beginning of this section, if the number of allowed bits per

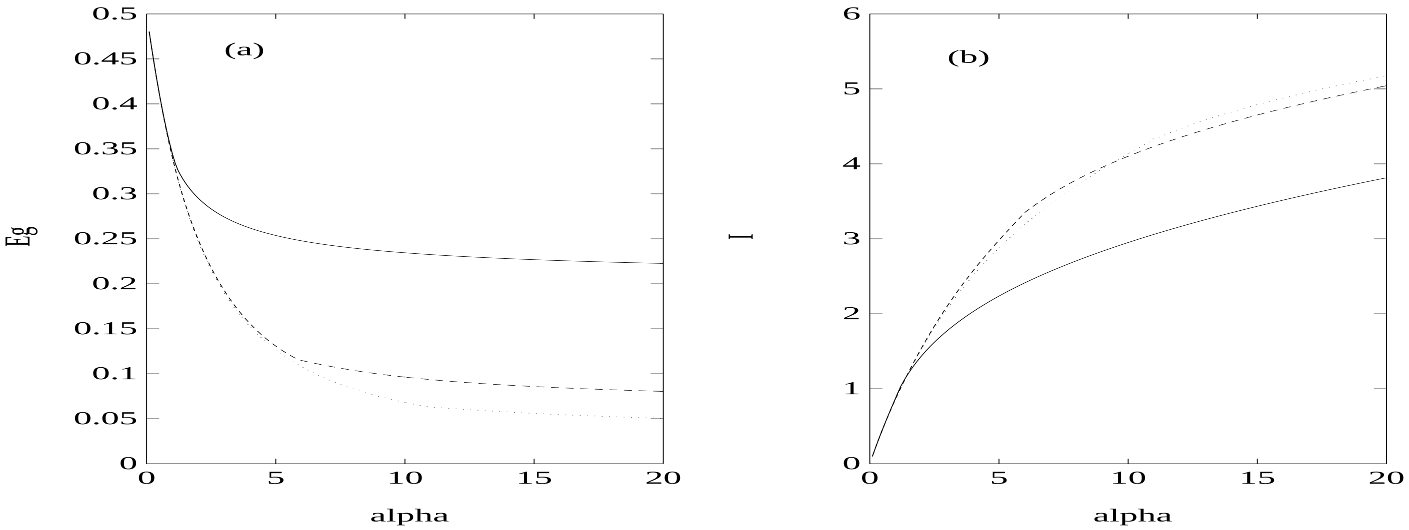

Figure 2: (a) Generalization error for the (L_s, R) model for $L_s = 1$ (solid line), $L_s = 5$ (dashes), $L_s = 10$ (dots). (b) Same as (a) but for the stochastic complexity.

weight is permitted to increase with the sample size, then we find that $\log \alpha$ bits per weight are sufficient, a large decrease compared to α.

In figure (2b) we plot the stochastic complexity for the different models as a function of sample size α. For small α we find that the shortest description length is provided by the $L_s = 10$ model. This is similar to the unrealizable $L_t = 5$ problem, where the shorter description length was obtained (for small α) for the model with the larger complexity. As α increases we again find that the simplest model, $L_s = 1$, possesses smallest description length while having the largest generalization error . Thus there is a regime of α where the description length criterion and the generalization error give opposite predictions concerning model selection. For large α of course they both agree as expected. As in the discrete teacher case we find a monotonic relationship between the generalization error and the stochastic complexity.

5 Discussion

We have investigated a class of simple neural network models in an attempt to shed light on the issues of data compression and prediction as well as model selection. In particular we have studied the stochastic complexity, which represents the shortest description length of the data using a given model h. We have found that the data can be greatly compressed even in the case where the model is insufficient to perfectly implement the data. While the behavior of the MSC model selection criterion is well understood (and indeed optimal) for large sample sizes, we have found great variability for finite sample sizes. It would thus seem fair to say that without further prior knowledge, the problem of model selection is insoluble for finite sample sizes. This result is perhaps not surprising, since there can be inifinitely many models giving rise to a bounded set of input/output pairs. Even within the limited domain of the hypothesis set $\mathcal{H}$ we find that although the optimal model is asymptotically approached, the convergence may be very slow.

Within the context of a single neural network architecture (i.e. given precision constraints) we have found that the relationship between the generalization error and the stochastic complexity is monotonic. We have not been able to derive an exact expression for the relationship and indeed it seems to depend sensitively on the details of the problem.

References

[Gar88] Gardner E. The space of interactions in neural network models. *J. Phys. A* 21:257-270.

[GD88] Gardner E. and B. Derrida. Optimal storage properties of neural networks, *J. Phys. A* **21** 271-284.

[FM93] The statistical mechanics of the Ising perceptron. *J. Phys. A* 1993, to appear.

[Kol65] Kolmogorov A.N., Three approaches to the quantitative definition of information. *Problems of Information Transmission*, **1**:4-7.

[LTS90] Levin E., N. Tishby and S. Solla. A statistical approach to learning and generalization in layered neural networks. *Proc. IEEE*, **78**:1568-1574.

[MF92] Meir R. & J.F. Fontanari. Learning from examples in weight constrained neural networks. *J. Phys. A*, 25: 1149-1168.

[MF93] Meir R. & J.F. Fontanari. On the precision constraints of threshold elements, submitted to NETWORK.

[Ris86] Rissanen J. Stochastic complexity and modeling, *Ann. Stat.* **14**(3): 1080-1100.

[Sha48] Shannon C.E. A mathematical theory of communication. *Bell. Syst. Tech. Journal*, **27**:379-423, 623-656.

[SST92] Seung S., H. Sompolinsky and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev. A***45**: 6056-6091.