Learning Sparse Networks From Poor Data

Goele Hollanders . @ .
Geert Jan Bex . @ .
Marc Gyssens . @ .

Department of Mathematics, Physics, and Computer Science, Hasselt University and Transnational University of Limburg, Hasselt, Belgium

Ronald L. Westra W @MICC. .
Karl Tuyls K.T @MICC. .

Department of Mathematics and Computer Science, Maastricht University and Transnational University of Limburg, Maastricht, The Netherlands

Abstract

This paper is concerned with the learning process of a sparse interaction network, for example, a gene-protein interaction network. The advantage of the process we purpose is that there will always be a student S that fits the teacher T very well with a relatively small data set and a high number of unknown components, i.e., when the number of measurements M is significantly smaller than the system size N.

To measure the efficiency of this learning process, we use the generalization error, ϵ_{gen}, which represents the probability that the student is a good fit to the teacher. From our experiments it follows that the quality of the fit depends on several factors: First, the ratio $\alpha = M/N$ of the number of measurements to the system size has a strong impact. Surprisingly, we find that a sudden identification transition occurs for value $\alpha \approx \alpha_{gen}$ which corresponds to $\epsilon_{gen} = 1/2$. From this sample size onwards the student will be a good fit to the teacher. Interestingly, the generalization threshold α_{gen}, will always be significantly smaller then 1. Second, the quality of the fit depends on the sparsity of the network. If the number of non-zero components increases, as sparsity disappears, the efficiency of the process will gradually increase. Finally there is an impact of the noise level. The learning process is robust to noise upto a certain threshold. We see that, at this level, the impact on the noise suddenly and dramatically increases as a consequence of which the student will no longer be a good fit to the teacher.

Keywords: machine learning, sparse systems, network reconstruction, robust identification.

1. Introduction and motivation

Over the last two decades research in the area of genetics and bio-informatics has increased spectacularly. One of the main contributing factors has been the development of microarray technologies, which has enabled the measurement of gene expression levels and profiles on a genome-wide scale.

In previous work (Westra et al., 2006) we were concerned with the identification and reconstruction of dynamic gene-protein interaction networks with intrinsic and extrinsic noise from empirical data, such as a set of microarray time series. To model the interactions amongst genes and proteins, we considered them as biochemical reactions and thus we represented them as rate equations.

$$\dot{x}_i(t) = \sum_{j=1}^{N} a_{ij} x_j(t) + \sum_{p=1}^{P} b_{ip} u_p(t) + \xi_i(t) - \lambda_i x_i(t) \qquad (1)$$

for $(1 \leq i \leq N)$ with

- a_{ij} the influence rate from gene j on gene i
- $x_j(t)$ the expression of gene j at time t
- b_{ip} the external stimuli rate from input p on gene i
- $u_p(t)$ the expression of external stimuli from input p at time t

Proceedings of the 18th Benelearn
P. Adriaans, M. van Someren, S. Katrenko (eds.)

- $\xi_i(t)$ any stochastic uncertainty for gene i at time t
- λ_i the decay rate

In many engineering applications, the number of measurements M available for system identification (also known as reverse engineering) and model validation is usually much larger than the system order N, which represents the number of genes. In the application of Westra et al. (Westra et al., 2006), though, the number of measurements M is typically much smaller than N because we have to deal with *poor data*. In practice, because of the high costs, there are a few microarray experiments and lots of genes. This substantial lack can give rise to an identifiability problem, in which case a larger subset of the model class is entirely consistent with the observed data and no unique model results. Since conventional techniques for system identification are not well suited to deal with such situations, it thus becomes important to work around this by exploiting as much additional information as possible about the underlying system, in particular the relation between the number of measurements and the number of genes, the sparsity of the gene regulatory network and the influence of noise.

The model constructed in (Westra et al., 2006) is not only capable of reconstructing gene-protein interaction networks, but can also reconstruct other networks based on a small number of available measurements.
Alternatively, the problem may be rephrased as the identification of the subset of features used by the teacher to generate its response. Since the teacher is sparse, only a subset of the available input values is actually processed. Hence, we can view this problem in the context of feature selection.
Because of the different possibilities we want to generalize this process of reconstruction and reformulate it as a learning process. A learning process is defined by Mitchell (Mitchell, 1997) as searching through a very large space of possible hypotheses to determine one that best fits the observed data and any prior knowledge held by the learner. Learning is useful when, first, large databases may contain valuable implicit regularities that can be discovered automatically, second, in poorly understood domains where humans might not have the knowledge needed to develop effective algorithms and, third, in domains where the program must dynamically adapt to changing conditions.
For our model we have a teacher, $T = (A_T, B_T)$, with $A_T = \mathbb{R}^{NxN}$ and $B_T = \mathbb{R}^{NxP}$. Each component of $A_T \in [-0.9, -0.1[\cup]0.1, 0.9]$ and each component of $B_T \in [-1, 1]$ and are uniform distributed. This teacher T is a matrix that represents the mRNA concentration per gene per measurement and has the knowledge about the number of genes (in general the number of rows of A_T), the number of external influences (in general the number of columns of B_T) and the sparsity, k_T. k_T represents the number of non-zero values per row of A_T and B_T together. For each measurement, m, the following equation holds:

$$\dot{x}_m = a_T x_m + b_T u_m \tag{2}$$

The aim is to reproduce the teacher's output for any input perfectly after seeing M examples, so we need a student, $S = (A_S, B_S)$ with $A_S = \mathbb{R}^{NxN}$ and $B_S = \mathbb{R}^{NxP}$, to learn. Learning means that the student has to determine the subset of features used by the teacher and in this context thus has to search for the interaction matrix, (A_S, B_S), that is an acceptable fit for the original one, (A_T, B_T), using the rate equation (2) and $(X, U, \dot{X})$. With an 'acceptable' fit we mean that the estimated matrix can differ from the original one with a small but acceptable error.

2. The learning process

The rate equation, which will be used to reconstruct interaction matrices and is described in Section 1 can be rewritten as a state-space system of the form:

$$\dot{X} = AX + BU + \xi \tag{3}$$

where

- $\dot{X} = (\dot{x}_1, \ldots, \dot{x}_m) \in \mathbb{R}^{NxM}$, the estimates of the state derivatives at $t_1, \ldots, t_M$
- $X = (x_1, \ldots, x_1) \in \mathbb{R}^{NxM}$, the full state observations at $t_1, \ldots, t_M$
- $U = (u_1, \ldots, u_m) \in \mathbb{R}^{PxM}$, the external stimuli at $t_1, \ldots, t_M$
- $\xi \in \mathbb{R}^{NxM}$, stochastic Gaussian white noise

are all known quantities.

It should be noted that this matrix equation can be treathed in a row-by-row fashion. Denoting the (unknown) i-th row of A_S by the row-vector α_i, the (unknown) i-th row of B_S by the row-vector β_i and the (known) i-th row of $\dot{X}$ by the row-vector δ_i. This yields the following decoupled set of N linear systems of equations of size MxN (with M the number of equations and N the number of unknowns):

$$\begin{bmatrix} \alpha_i & \beta_i \end{bmatrix} \begin{bmatrix} X \\ U \end{bmatrix} = \delta_i \ , \quad (1 \le i \le N) \tag{4}$$

The process to learn these interaction matrices can be formulated as follows:

In each of these linear systems of equations, the N-vector α_i and the P-vector β_i are to be computed from M equations, with M typically smaller than N, to deal with the poor data property. This means that there are too many degrees of freedom because there are a small number of equations, M, and a lot of unknown variables, N. To deal with this problem, an extra condition is imposed. More precisely a sparsity constraint on α_i and β_i is added, which dictates that a number of components of α_i and β_i have to be zero. Some studies about computing a sparse solution to a consistent underdetermined linear system of equations $Cx = D$ has been conducted. The L_0-minimization maximizes the number of zeros, but is too complex to compute. Therefore L_1-minimization is used because it gives, under certain conditions, a good approximation as noted by Fuchs (Fuchs, 2003; Fuchs, 2004). For more details about this technique see also (Peeters & Westra, 2004) and (Westra, 2005). It is well known that this problem can be reformulated as an LP-problem (linear programming problem). In other words, finding a vector having as many zeros as possible can be replaced by the much simpler problem of finding a vector for which $\|S\|_1$ is minimal.
For simplicity of representation, we incorporate matrix B_S in matrix A_S ($dim(A_S) = N \times (N + P)$) and matrix U in matrix X ($dim(X) = (N + P) \times M$). So the LP problem can be reformulated by: minimize the L_1-norm of A_S

$$\min_{A_S} \|A_S\|_1 \iff \min_{a_{S_{i,1}} \ldots a_{S_{i,N+P}}} \sum_{j=1}^{N+P} |a_{S_{i,j}}|\,, \quad (1 \le i \le N) \quad (5)$$

subject to the M conditions

$$\dot{x}_m = a_S\, x_m \quad (6)$$

which represents the training set $\chi = (X_m, \dot{X}_m)$ for $m = 1 \ldots M$. LP yields a student S that always faithfully reproduces the training set, but may or may not be equal to the teacher T. In the context of learning the latter implies that the student has memorized or stored the examples in the training set perfectly, but was unable to generalize beyond that point. This fact will be taken into account by introducing the generalization error. It represents the failure of the student to reproduce the teacher's output given a training set of size M. For more details see Section 3.

The main question of this paper is the following: how many measurements, $M_{gen} = \alpha_{gen} N$, does the student need to be supplied with in order to determine the subset of features used by the teacher? First, it is important to determine the required size of the training set for a student to successfully learn the teacher's character. If the number of available measurements is too small as a function of the number of rows, the learning process is not capable to construct a good fit. Secondly, the sparsity k also has a strong influence on M and thirdly, the quality of the result also depends on the noise level, ξ. We try to establish this by experiments. These experiments are presented in Section 3.

3. Experiments

The success of the learning process described in Section 2 depends on the relations between the parameter values. To investigate these, several numerical experiments have been conducted.

All experiments were performed on a PC with an Intel Pentium M processor of 1.73 GHz and 1 GB RAM memory under Windows XP Professional, using Matlab 6.5 Release 13 including the Optimization Toolbox. The latter's routine `linprog` was used to solve LP problems; its default solution method is a primal-dual interior point method, but an active set method can optionally be used, too. For larger problems, it turned out to be essential for obtaining reasonable computation times that the LP problems were solved by application of the active set method on the dual problem formulation. Therefore, this method was adopted throughout all the experiments.

Since results can depend on the particularities of given data and the original system that generated it, all experiments have been performed on a number of independent runs on randomly selected data and systems. Hence they convey the behavior of our approach "on average".

In line with the definitions above, we use the parameters N and M to quantify the size and complexity of the input. To generalize these input-parameters we define $\alpha = M/N$. In addition, the sparsity of the interaction matrix A_T is measured by the number of non-zero entries per row and denoted by k_T (which should be much smaller than N). Also for this parameter we use a generalization $\kappa_T = k_T/N$. To complete the system's data set, some stochastic Gaussian white noise ξ is added to the input data set. It is normally distributed with zero mean and some standard deviation σ that determines the noise level. To quantify the quality of the fit, the resulting approximation A_S, a performance measure is introduced: $\overrightarrow{S} \circ \overrightarrow{T}$.

As already described in Section 1, for the quality of the learning process it is important to evaluate the training set χ. The number of available measurements M depends on the number of rows N because if there are not enough measurements, the student is not a good fit to the teacher. This phenomenon is shown in Figure 1. Results have been obtained by averaging over 50 independently and randomly selected instances of teacher and training set for three different system sizes $N = 100, 160, 300$, each with sparsity $\kappa_T \approx 0.031$. As mentioned before, by the use of LP we al-

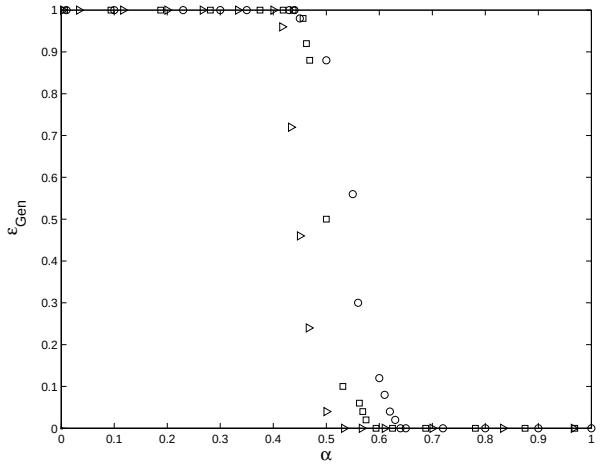

Figure 1. The probability of an error ϵ_{gen} as a function of learning the training set of size $\alpha = M/N$ for $N = 100$(circles), $N = 160$(squares) and $N = 300$(triangles) for constant $\kappa_T \approx 0.031$ and for a full matrix A_T.

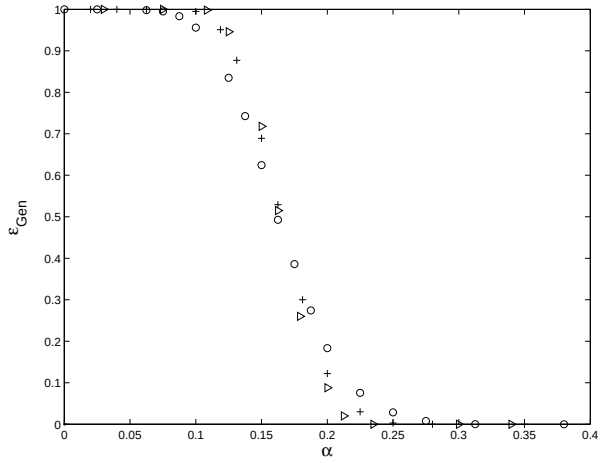

Figure 2. The probability of an error ϵ_{gen} as a function of learning the training set of size $\alpha = M/N$ for $N = 100$(diamonds), $N = 160$(squares) and $N = 300$(triangles) for constant $\kappa_T \approx 0.031$ and for only one row A_{T_n}.

ways get a student S that is capable to reproduce the training set correctly, but there is a possibility that this student may not be equal to the teacher T. In the context of learning this situation implies that the student has memorized or stored the examples in the training set perfectly, but was unable to generalize beyond that point. For small values of α, the algorithm fails to reproduce the teacher with probability $\epsilon_{gen} \approx 1$. However, for increasing size of the training set, one notes that LP generates a student that equals the teacher with probability $1 - \epsilon_{gen} \approx 1$. The transition from one regime to the other is quite sudden, especially for increasing system size N. That is the reason why the generalization threshold is introduced. It represents the minimal size of the training set from which the student is able to generalize. Since the size of the training set necessary for generalization depends on a specific instance of the teacher and the training set, the generalization threshold can be defined more rigorously as the fraction $\alpha_{gen} = M_{gen}/N$ so that the student and the teacher will be equal after training with probability 1/2 for a large number of independently and randomly selected instances of T and χ. Figure 1 shows that, if α is sufficiently large, i.e., if $\alpha \geq \alpha_{gen}$, the student S fits the teacher very well. Furthermore we notice that $\alpha_{gen} < 1$.

Moreover, Figure 1 illustrates that α_{gen} depends on the size of the system N for constant κ_T because the transition towards generalization occurs for smaller values of α as the system size N increases.

If only one equality (one row of A_{T_n}) is considered, as in Figure 2, we see that the transition towards generalization occurs increasingly sudden for larger systems N. Furthermore we see that α_{gen} is independent of N for constant κ_T.

The relation between the generalization error curves for a system of N rows in Figure 1 and those for one with a single row in Figure 2 can be understood as follows. The generalization threshold α_{gen} for the former is—by definition—reached when $\varepsilon_{\text{gen}} = 1/2$. However, this implies that for half the number of independent runs, *all* rows must be identified correctly. This means that the acceptable error rate $\varepsilon_{\text{acc}} \ll 1/2$ for each row, even for very small systems. More precisely, the acceptable generalization error per row can be computed from $1/2 = (1-\epsilon_{acc})^N$, or, $\epsilon_{acc} = 1-(1/2)^{1/N}$. The size of the training set required to attain the error rate ϵ_{acc} is denoted by α_{acc} and can—at least in principle—be read from the curves in Figure 2 for the desired value of the system size N. By definition, $\alpha_{\text{acc}} = \alpha_{\text{gen}}$ which elucidates the relation between Figures 1 and 2. In Figure 2, one observes that for increasing system size, the system identification transition is increasingly sudden, i.e., the derivate at $\epsilon_{gen} = 1/2$ increases with N. On the other hand, ϵ_{acc} decreases with N, and hence it is not a priori clear from Figure 2 how α_{gen} will vary as a function of N. Figure 3 allows to derives an upper bound for α_{acc} and hence a lower bound

for the suddeness of the transition in Figure 2 for values of $N \to \infty$. More precisely, $0.17 \leq \alpha_{acc} \leq 0.39$, with the lower bound, 0.17, the size of the training set necessary to learn just one row (see Figure 2) and 0.39 the upper bound determined by interpolation.

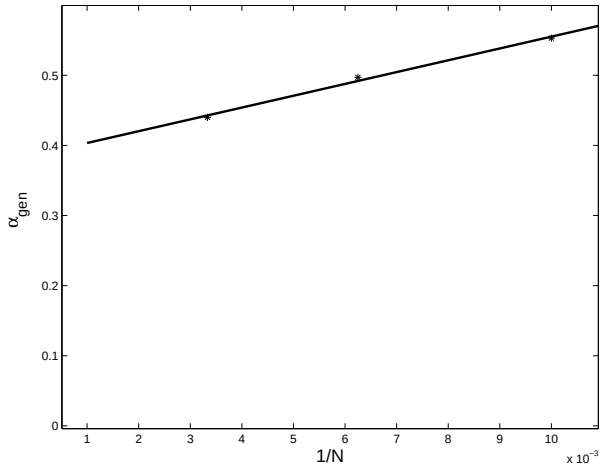

Figure 3. The training set of size $\alpha_{gen} = M_{gen}/N$ as a function of $1/N$, the blue stars are measured from figure 1, the red line is an approximation based on these stars.

Note that the probabilistic approach outlined above also explains another qualitative difference between Figures 1 and 2. While the curves in Figure 2 are anti-symmetric with respect to $\varepsilon_{\text{gen}} = 1/2$, this is not the case in Figure 1. Since a failure to identify even a single row of the matrix results in the failure of the system as a whole, the system identification transition sets in for values of $\alpha_{\text{gen}} - \alpha \approx 0$, but extends to—comparared to Figure 2—larger values of $\alpha - \alpha_{\text{gen}}$.

This result is rather surprising since one would expect that N examples would be needed to determine N unknown values. However, it is clear that $\kappa_T \ll \alpha_{gen} < 1$, which shows that less information than naively expected is necessary to achieve good generalization. This is a consequence of the sparsity of the teacher. It turns out to be much easier to identify the input features used by a sparse teacher to determine its output than to determine the interactions at a non-sparse teacher. Obviously one needs more than $k_T = \kappa_T N$ examples to determine k_T non-zero values since their positions also need to be identified.

Figure 4 shows the student-teacher overlap $\vec{S} \circ \vec{T}$ as a function of the training set size. $\vec{S} \circ \vec{T}$ represents the correctness (=1-error) of the results of the learning processes. The error is defined as follows: each value of the student $A_{S_{n,n}}$ will be compared with the corresponding value of the teacher $A_{T_{n,n}}$. The error is the sum of the false positives and false negatives in the student. If $A_{S_{n,n}}$ is non-zero while it should be zero because $A_{T_{n,n}}$ is zero, there is a false positive and if $A_{S_{n,n}}$ is zero while it should be non-zero or when $A_{S_{n,n}}$ has a different sign than $A_{T_{n,n}}$ there is a false negative.

Again, results have been obtained by averaging over 50 independently and randomly selected instances of teacher and training set for three different system sizes $N = 100, 160, 300$, each with sparsity $\kappa_T \approx 0.031$. Figure 5 is also a representation of the quality of the learning as a function of α, but only for one row/equation of A_S and A_T and only one run of a randomly selected instance of the teacher and the training set for one system size $N = 240$ and for a constant sparsity $\kappa_T \approx 0.037$. Both figures show

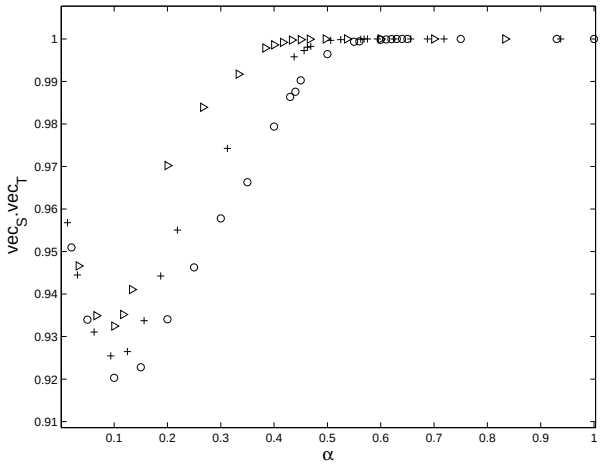

Figure 4. The quality of the learning process as a function of the training set of size $\alpha = M/N$ for $N = 100$(circles), $N = 160$(squares) and $N = 300$(triangles) and for constant $\kappa_T \approx 0.031$.

an initial decrease, followed by an increase to $\vec{S} \circ \vec{T} = 1$. Observe that the startpoints, $M = 1$ for all system sizes N, are equal to approximately $(1 - \kappa_T)$. The initial decrease is because the number of non-zero components is very low and increases as a function of N until a certain point. At this point the identification of the system starts and there is an increase until the number of necessary measurements is available. From this point onwards the student fits the teacher perfectly and $\vec{S} \circ \vec{T} = 1$. As in Figure 1, there is an increasingly sudden transition from one regime to the other for higher values for N. Also the minimum value of $\vec{S} \circ \vec{T}$ decreases in Figure 4 for increasing system sizes. These sudden transitions are more clear in Figure 5 (represents just one typical run), because Figure 4 has the same results but has to take all rows of A_S into account and is averaged over 50 runs.

Until now we have used a fixed value for the sparsity of

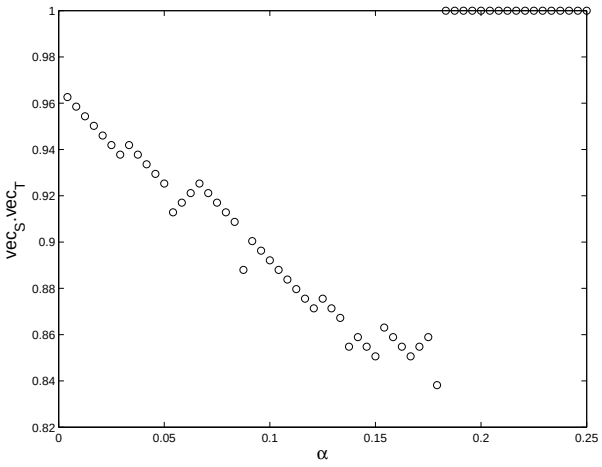

Figure 5. The quality of the learning process as a function of the training set of size $\alpha = M/N$ for only one row, with $N = 240$ and a constant $\kappa_T \approx 0.037$.

the teacher, κ_T, but the sparsity has also an impact on the learning process. Results have been obtained again by averaging over 50 independently and randomly selected instances of teacher and training set for two different system sizes $N = 100, 160$. Figure 6 shows that if κ_T increases α_{gen} increases too. It is obvious that the more non-zero values there are per row of A_T, the more measurements are necessary to fit this teacher A_T. Furthermore, it is also important

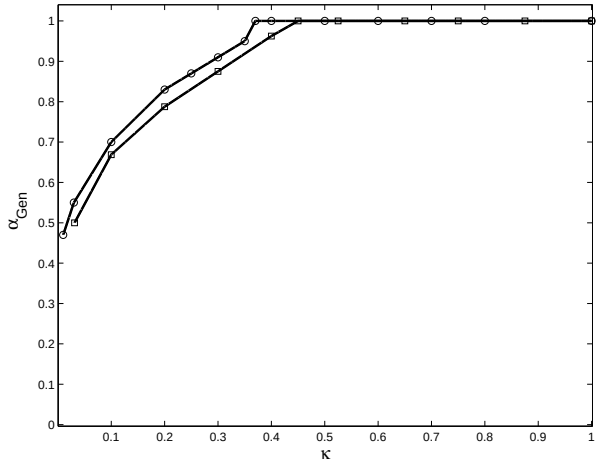

Figure 6. The training set of size $\alpha_{gen} = M_{gen}/N$ as a function of the sparsity $\kappa_T = k_T/N$ for $N = 100$(circles) and $N = 160$(squares).

to observe that there exists an upper bound for the degree of sparsity, for example, the upper bound is approximately at $\kappa_T \approx 0.37$ for $N = 100$. We also can conluce that this upper bound is dependent on the system size N. This upper bound means that if κ_T is larger than this upper bound, the number of measurements has to be maximal so that $\alpha_{gen} = 1$ for realistic system sizes.

Finally, as mentioned, it is also important to observe the influence of noise on the learning process. Therefore we added to the input data set some stochastic Gaussian white noise ξ with zero mean.The standard deviation σ of this noise distribution can be interpreted as the noise level. Results have been obtained by averaging over 50 independently and randomly selected instances of teacher and training set for one system size $N = 100$, with sparsity $\kappa_T \approx 0.031$. As we can expect, α_{gen} increases when σ in-

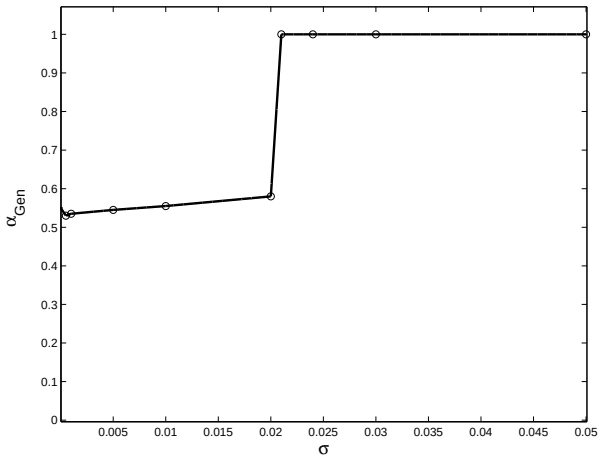

Figure 7. The training set of size $\alpha_{gen} = M_{gen}/N$ as a function of the noise level σ.

creases: we see that the increase varies as a function of σ. More precisely, the increase of α_{gen} is more or less linear upto $\sigma \approx 0.021$. From that point onwards $\alpha_{gen} = 1$. This value of σ is the threshold for the noise level beyond which the model is not able to learn with less than N examples. Further more, we known that the standard deviation of the system's response $\dot{X}$, denoted by $\sigma_{\dot{X}}$, is of order 1. So we conclude that the learning process is robust to noise upto a noise level of 2 %.

4. Conclusions

Using experiments, we have addressed answers to the following questions:

- How many measurements $M_{gen} = \alpha_{gen}N$ does the student S need to fit the teacher T?
- How does the sparsity influence α_{gen}?
- How does the noise level influence α_{gen}?

From our experiments, we can conclude that the student fits the teacher very well when α is significantly smaller than

1 on two conditions. First, there is an upper bound for the sparsity which depends on the system size N. Second, the noise level should not exceet a threshold.

In summary, our experiments clearly demonstrate that for high degrees of sparsity and relatively moderate noise levels, the student already fits the teacher with great accuracy for a small sample size.

References

de Jong, H. (2002). Modeling and simulation of genetic regulatorysystems: A literature review. *Computational Biology*, *9*, 67–103.

Fuchs, J. (2003). More on sparse representations in arbitrary bases. *Proc. 13th IFAC Symp. on System Identification*, 13571362.

Fuchs, J. (2004). On sparse representations in arbitrary redundant bases. *IEEE Trans. on IT*.

Glass, L., & Kauffman, S. (1973). The logical analysis of continuous non-linear biochemical control networks. *J.Theor.Biol.*, *39(1)*, 103–129.

Goldbeter, A. (2002). Computational approaches to cellular rhythms. *Nature*, *420*, 238–45.

Gouze, J., Hernandez, C., Page, M., Sari, T., & Geiselmann, J. (2004). Qualitative simulation of genetic regulatory networks usingpiecewise-linear models. *Bull Math Biol.*, 301–40.

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.

Novak, B., & Tyson, J. (1997). Modeling the control of dna replication in fission yeast. *PNAS*, *94*, 9147–9152.

Peeters, R. L. M., & Westra, R. L. (2004). On the identification of sparse gene regulatory networks. *Proc. of the 16th Intern. Symp. on Mathematical Theory of Networks and Systems (MTNS2004)*.

Westra, R. L. (2005). Piecewise linear dynamic modeling and identification of gene-protein interaction networks. *Nisis/JCBWorkshop reverse engineering*.

Westra, R. L., Hollanders, G., Bex, G., Gyssens, M., & Tuyls, K. (2006). The identification of dynamic gene-protein networks.

Yeung, M. K. S., Tegnér, J., & Collins, J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Nat. Acad. Science*, *99(9)*, 6163–6168.