Advances in Large Margin Classifiers

edited by
Alexander J. Smola
Peter Bartlett
Bernhard Schölkopf
Dale Schuurmans

The MIT Press
Cambridge, Massachusetts
London, England

1 Support Vectors and Statistical Mechanics

Rainer Dietrich
Institut für Theoretische Physik
Julius-Maximilians-Universität
Am Hubland
D-97074 Würzburg, Germany
dietrich@physik.uni-wuerzburg.de
http://theorie.physik.uni-wuerzburg.de/∼dietrich

Manfred Opper
Department of Computer Science and Applied Mathematics
Aston University
Aston Triangle
Birmingham B4 7ET, UK
opperm@aston.ac.uk
http://www.ncrg.aston.ac.uk/People/opperm/Welcome.html

Haim Sompolinsky
Racah Institute of Physics and Center for Neural Computation
Hebrew University
Jerusalem 91904, Israel
haim@fiz.huji.ac.il

We apply methods of Statistical Mechanics to study the generalization performance of Support Vector Machines in large dataspaces.

1.1 Introduction

Many theoretical approaches for estimating the generalization ability of learning machines are based on general, distribution independent bounds. Since such bounds hold even for very unfavourable data generating mechanisms, it is not clear a priori how tight they are in less pessimistic cases.

Hence, it is important to study models of nontrivial learning problems for which we can get exact results for generalization errors and other properties of a trained learning machine. A method for constructing and analysing such learning situations has been provided by Statistical Mechanics. Statistical Mechanics is a field of Theoretical Physics which deals with a probabilistic description of complex systems that are composed of many interacting entities. Tools originally developped to study the properties of amorphous materials enable us to conduct controlled, *analytical* experiments for the performance of learning machines for specific types of data distributions when the numbers of tunable parameters and examples are large. While often statistical theories provide asymptotic results for sizes of the training data sample that are much larger than some intrinsic complexity of a learning machine, in contrast, the so called 'thermodynamic limit' of Statistical Mechanics allows to simulate the effects of small *relative* sample sizes. This is achieved by taking the limit where both the sample size and the number of parameters approaches infinity, but an appropriate ratio is kept fixed.

Starting with the pioneering work of Elizabeth (4) this approach has been successfully applied during the last decade to a variety of problems in the context of neural networks (for a review, see e.g. (9; 13; 7)). This chapter will deal with an application to learning with Support Vector Machines (SVMs). A somewhat more detailed analysis which was designed for readers with a Statistical Physics background, can be found in (2).

1.2 The basic SVM setting

We will restrict ourselves to SVM classifiers. They are defined (for more explanations, see the introductory chapter to this book) by a nonlinear mapping $\Phi(\cdot)$ from input vectors $\mathbf{x} \in \mathbb{R}^N$ into a feature space $\mathcal{F}$. The mapping is constructed from the eigenvectors $\psi_j(\mathbf{x})$ and eigenvalues λ_j of an SVM kernel $k(\mathbf{x}, \mathbf{y})$ via $\Phi(\mathbf{x}) = (\sqrt{\lambda_1}\psi_1(\mathbf{x}), \sqrt{\lambda_2}\psi_2(\mathbf{x}), \ldots)$.

The output y of the SVM can be represented as a linear classification

$$\operatorname{sgn}\left(\Phi(\mathbf{x}) \cdot \mathbf{w}\right) = \operatorname{sgn}\left(\sum_{j=1}^{N_{\mathcal{F}}} \sqrt{\lambda_j}\psi_j(\mathbf{x}) w_j\right) \tag{1.1}$$

in feature space, where for simplicity, we have set the bias term equal to zero. For a realizable setting, the weights w_j, $j = 1, \ldots, N_{\mathcal{F}}$ are adjusted to a set of example pairs $\{(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)\}$ by minimizing the quadratic function $\frac{1}{2}||\mathbf{w}||^2$ under the constraints that $y(\Phi(\mathbf{x}) \cdot \mathbf{w}) \geq 1$ for all examples.

1.3 The learning problem

teacher-student framework

We assume a simple noise free scenario, where the generation of data is modelled within the so called teacher-student framework. Here, it is assumed that some classifier (the teacher) which has a similar representation as the machine of interest, gives the correct outputs to a set of randomly generated input data. The generalization error can be measured as the probability of disagreement on a random input between teacher and student machine. In our case, we choose the representation

$$y_i = \operatorname{sgn}\left(\sum_j \sqrt{\lambda_j} B_j \psi_j(\mathbf{x}_i)\right). \tag{1.2}$$

All nonzero components are assumed to be chosen independently at random from a distribution with zero mean and unit variance. We will also consider the case, where a finite fraction of the B_j are 0 in order to tune the complexity of the rule. Finally, the inputs $\mathbf{x}_i$ are taken as independent random vectors with a uniform probability distribution $D(\mathbf{x})$ on the hypercube $\{-1,1\}^N$. We are interested in the performance of the SVM averaged over these distributions.

eigenvalue decomposition

We will specialize on a family of kernels which have the form $k(\mathbf{x},\mathbf{y}) = K\left(\frac{\mathbf{x}\cdot\mathbf{y}}{N}\right)$, where, for simplicity, we set $K(0) = 0$. These kernels are permutation symmetric in the components of the input vectors and contain the simple perceptron margin classifier as a special case, when $K(z) = z$. For binary input vectors $\mathbf{x} \in \{-1,1\}^N$, the eigenvalue decomposition for this type of kernels is known (5). The eigenfunctions are products of components of the input vectors, i.e. $\psi_i(\mathbf{x}) = 2^{-N/2}\prod_{j\in S_i} x_j$, which are simple monomials, where $S_i \subseteq \{1,\ldots,N\}$ is a subset of the components of $\mathbf{x}$. For polynomial kernels, these features have also been derived in (11). The corresponding eigenvalues are found to be $\lambda_i = 2^{N/2}\sum_{\mathbf{x}} k(\mathbf{e},\mathbf{x})\psi_i(\mathbf{x})$, with $\mathbf{e} = (1,\ldots,1)^T$. They depend on the cardinality $|S_i|$ of the set S_i only. For $|S_i| = 1$, the eigenfunctions are the N linear functions x_j, $j = 1,\ldots,N$. For $|S_i| = 2$, we have the $N(N-1)/2$ bilinear combinations $x_i x_j$ etc. The behaviour of the eigenvalues for large input dimension N is given by $\lambda_i \simeq \frac{2^N}{N^{|S_i|}} K^{(|S_i|)}(0)$. $K^{(l)}$ denotes the l-th derivative of the function K. The rapid decrease of the eigenvalues with the cardinality $|S_i|$ is counterbalanced by the strong increase of their degeneracy which grows like $n_{|S_i|} = \binom{N}{|S_i|} \simeq N^{|S_i|}/|S_i|!$. This keeps the overall contribution of eigenvalues $\sum_{|S_i|=l}\lambda_i n_{|S_i|}$ for different cardinalities l of the same order.

1.4 The approach of Statistical Mechanics

The basic idea to map SVM learning to a problem in Statistical Mechanics is to define a (Gibbs) measure $p_\beta(\mathbf{w})$ over the weights $\mathbf{w}$ which in a specific limit is

concentrated at the weights of the trained SVM. This is done by setting

$$p_\beta(\mathbf{w}) = \frac{1}{Z} e^{-\frac{1}{2}\beta||\mathbf{w}||^2} \prod_{i=1}^{m} \Theta\left(y_i \sum_{j=1}^{N_\mathcal{F}} \sqrt{\lambda_j}\psi_j(\mathbf{x}_i) w_j - 1 \right). \tag{1.3}$$

$\Theta(x)$ is the unit step function which equals 1 for $x \geq 0$ and 0 else. Z normalizes the distribution. In the limit $\beta \to \infty$, this distribution is concentrated at the minimum of $||\mathbf{w}||^2$ in the subspace of weights where all arguments of the Θ functions are non-negative. This is equivalent to the conditions of the SVM quadratic programming problem. A different approach has been discussed in (3), where the Kuhn Tucker conditions of the optimization problem have been directly implemented into a Statistical Mechanics framework. It will be interesting to see, if this method can also be applied to the generalization problem of SVMs.

The strategy of the Statistical Mechanics approach consists of calculating expectations of interesting quantities which are functions of the weight vector $\mathbf{w}$ over both the distribution (1.3) and over the distribution of the training data. At the end of the calculation, the limit $\beta \to \infty$ is taken. These averaging procedures can be performed analytically only in the limit where $N \to \infty$ and $m \to \infty$. They require a variety of delicate and nontrivial manipulations which for lack of space cannot be explained in this contribution. One of these techniques is to apply a central limit theorem (valid in the 'thermodynamic limit') for carrying out expectations over the random inputs, utilizing the fact that the features ψ_j are orthogonal with respect to the chosen input distribution. This is the main reason, why we prefer to work in high-dimensional feature space rather than using the low dimensional kernel representation. A review of the standard techniques used in the Statistical Mechanics approach and their application to the generalization performance of neural networks can be found e.g. in (9; 13; 7)), a general review of the basic principles is (6).

thermodynamic limit

The results of our analysis will depend on the way, in which the two limits $N \to \infty$ and $m \to \infty$ are carried out. In general, one expects that a decay of the generalization error ϵ_g to zero should occur only when $m = \mathcal{O}(N_\mathcal{F})$, because $N_\mathcal{F}$ is the number of parameters of the data model. Nevertheless, when the mapping Φ contains a reasonably strong linear part, ϵ_g may drop to small values already on a scale of $m = \alpha N$ examples. Hence, in taking the limit $N \to \infty$, we will make the general ansatz $m = \alpha N^l$, $l \in \mathbb{N}$ and discuss different regions of the generalization performance by varying l. Our model differs from a previous Statistical Mechanics approach to SVMs (1) where the dimension of the feature space grew only linear with N.

1.5 Results I: General

One of the most basic and natural quantities which result from the calculation is a so called order parameter which for the SVM is defined by

$$R = \sum_i \Lambda_i \langle w_i B_i \rangle \tag{1.4}$$

where $\Lambda_i := \lambda_i / 2^N$, and $\langle ... \rangle$ denotes an average with respect to the distribution (1.3) and the distributions of the data and of the teacher vector. R is a weighted overlap between the teacher and SVM student weight vectors. This similarity measure between teacher and student allows us to express the generalization error by $\epsilon_g = \frac{1}{\pi} \arccos \frac{R}{\sqrt{Bq}}$. Here $B = \sum_i \Lambda_i \langle (B_i)^2 \rangle$ and $q_0 = \sum_i \Lambda_i \langle (w_i)^2 \rangle$ denote specific squared norms of the teacher and student weight vectors. Note that by the specific form of ϵ_g, the teacher's rule is perfectly learnt when the student vector points in the same direction as the teacher irrespectively of the student vector's length. Furthermore, an analysis of the contributions coming from eigenvectors of different complexities (i.e. cardinalities $|S_i|$) will give us an intuitive understanding of the SVMs inference of the rule.

generalization error

As a general result of our analysis, we find that if the number of examples is scaled as $m = \alpha N^l$,

scaling of number of inputs

- All high order components B_i are completely undetermined, i.e. $R^{(+)} := \sum_{|S_i|>l} \Lambda_i \langle w_i B_i \rangle \to 0$, and also that $q_0^{(+)} := \sum_{|S_i|>l} \Lambda_i \langle (w_i)^2 \rangle \to 0$, in the large N limit.

This does not mean that the values of the corresponding weights w_i are zero, they are just too small to contribute in the limit to the weighted sums (1.4).

- All low order components are *completely* determined, in the sense that $w_i = cB_i$ for all i with $|S_i| < l$, where c depends on α only. The only components which are actually learnt at a scale l are those for $|S_i| = l$.

To illustrate this behaviour for the simplest case, we study quadratic kernels of the form $K(x) = (1-d)x^2 + dx$, where the parameter d, $0 < d < 1$, controls the nonlinearity of the SVM's mapping. The eigenvectors of lowest complexity are just the N linear monomials $\sim x_j$, and the remaining ones are the $N(N-1)/2$ quadratic terms of the form $x_i x_j$. The learning curve is shown in Fig. 1.1, where we have included results from simulations for comparison.

If the number of examples scales linearly with the input dimension, i.e. $m = \alpha N$ (left side of Fig. 1.1), the SVM is able to learn only the linear part of the teacher's rule. However, since there is not enough information to infer the remaining $N(N-1)/2$ weights of the teacher's quadratic part, the generalization error of the SVM reaches a nonzero plateau as $\alpha \to \infty$ according to $\epsilon_g(\alpha) - \epsilon_g(\infty) \sim \alpha^{-1}$. The height of the plateau is given by $\epsilon_g(\infty) = \pi^{-1} \arccos(d)$, which increases from zero at $d = 1$, when the kernel is entirely linear, to $\epsilon_g = \frac{1}{2}$ at $d = 0$ when only quadratic features are present.

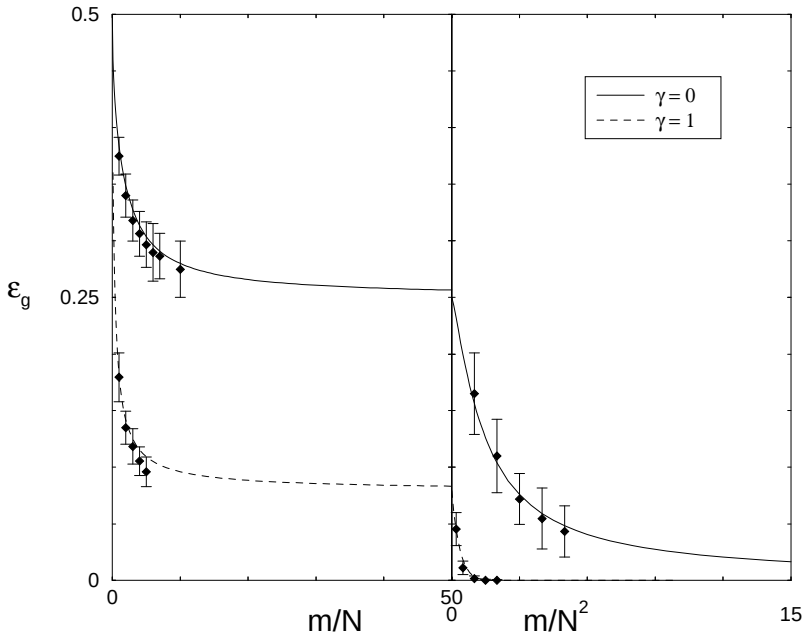

Figure 1.1 Decrease of the generalization error on different scales of examples, for quadratic SVM kernel learning a quadratic teacher rule ($d = 0.5$, $B = 1$) and various gaps γ. Simulations were performed with N= 201 and averaged over 50 runs (left and next figure), and N= 20, 40 runs (right).

If we increase the number of examples to grow quadratically with N, i.e. $m = \alpha N^2$ (right side of Fig. 1.1), the generalization error will decrease towards zero with a behavior $\sim 1/\alpha$ asymptotically, where the prefactor does not depend on d.

polynomial kernels

The retarded learning of the more complex components of the mapping Φ generalizes to kernels which are polynomials of higher order $z > 2$. On the scale of $m = \alpha N^l$ examples, when $l < z$, the generalization error decreases to a plateau as $\alpha \to \infty$ which is given by

$$\epsilon_g = \frac{1}{\pi} \arccos \sqrt{\frac{\sum_{j=1}^{l} \frac{K^{(j)}(0)}{j!}}{K(1)}}. \tag{1.5}$$

Only at the highest scale $m = \alpha N^z$, we get an asymptotical vanishing of the generalization error to zero as $\epsilon_g \approx \frac{0.500489}{z!}\alpha^{-1}$.

1.6 Results II: Overfitting

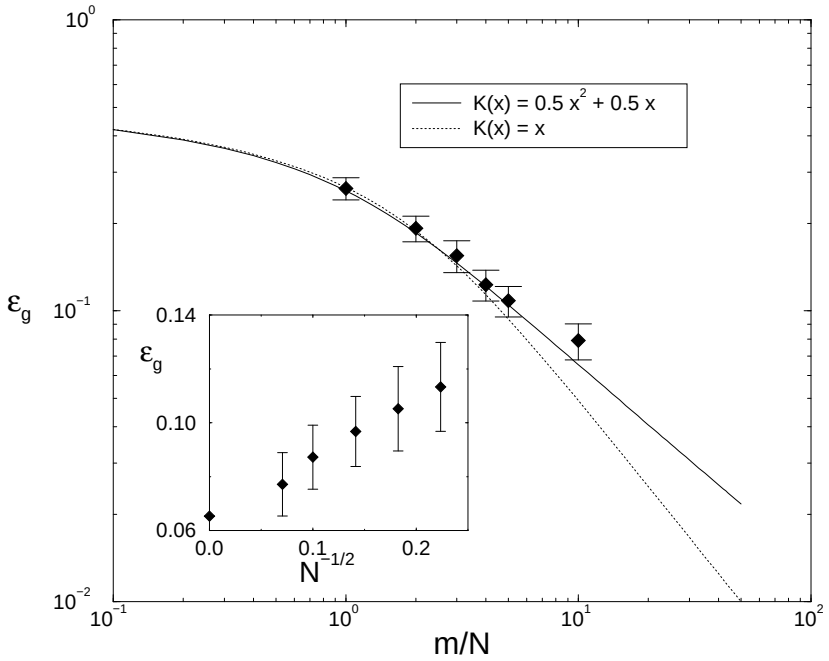

Figure 1.2 Learning curves for linear student and quadratic SVM kernels, all learning a linear teacher rule ($B = d$). For $\alpha = 10$, a finite size scaling is shown in the inset.

As the next problem, we study the ability of the SVM to cope with the problem of overfitting when learning a rule which has a much lower complexity than the mapping Φ. We model such a problem by keeping the SVM quadratic, but choosing a data generating mechanism which is defined by a simple *linear* separation of examples. This is achieved by setting $|B_i| = 1$ for $|S_i| = 1$ and $|B_i| = 0$ for the higher order components. Our results for the generalization error are shown in Fig. 1.2, where the number of examples is scaled as $m = \alpha N$. Surprisingly, although the complexity of the SVM is by far higher than the underlying rule, only a rather weak form of overfitting is observed. The SVM is able to learn the N teacher weights B_i on the correct scale of $m = \alpha N$ examples. The asymptotic rate of convergence is $\epsilon_g \sim \alpha^{-2/3}$. If we had used a simple linear SVM for the same task, we would have learned the underlying concept only slightly faster at the rate $\epsilon_g \sim \alpha^{-1}$.

We can compare these results with simple bounds on the expected generalization error as described in section ?? of the introductory chapter. E.g., the expectation of the ratio of the number of support vectors over the total number of examples m yields an upper bound on ϵ_g (12). Calculating the expected number of support vectors within the Statistical Mechanics approach yields an asymptotic decay $\sim \alpha^{-1/3}$ for this bound which decays at a slower rate than the actual ϵ_g.

1.7 Results III: Dependence on the input density

One can expect that if the density of inputs acts in a favourable way together with the teacher's concept, learning of the rule will be faster. We have modelled such a situation by constructing an input distribution which is *correlated* with the teacher weights B_i by having a gap of zero density of size 2γ around the teacher's decision boundary. In this case we expect to have a large margin between positive and negative examples. The density for this model is of the form $D(\mathbf{x}) \sim \Theta\left(|\sum_i \sqrt{\lambda_i} B_\rho \psi_i(\mathbf{x})| - \gamma\right)$.

For a quadratic SVM learning from a quadratic teacher rule, we observe a faster decay of the generalization error than in the case of a uniform density. However, on the linear scale $m = \alpha N$ (Fig. 1.1) the asymptotic decay is still of the form $\epsilon_g(\alpha) - \epsilon_g(\infty) \sim \alpha^{-1}$. A dramatic improvement is obtained on the highest scale $m = \alpha N^2$, where the generalization error drops to zero like $\epsilon_g \sim \alpha^{-3} e^{-\hat{c}(\gamma)\alpha^2}$. In this case, the mismatch between the true generalization error and the simple bound based on the fraction of support vectors is much more striking. The latter decreases much slower, i.e. only algebraically with α.

1.8 Discussion and Outlook

The present work analysed the performance of SV Machines by methods of Statistical Mechanics. These methods give distribution dependent results on generalization errors for certain simple distributions in the limit of high dimensional input spaces.

Why do we expect that this somewhat limited approach may be of interest to the machine learning community? Some of the phenomena discussed in this chapter could definitely be observed qualitatively in other, more general approaches which are based on rigorous bounds. E.g., the recently introduced concept of *luckiness* (10; 8) applied to the case of the favourable density with a gap would give smaller generalization errors than for a uniform density. This is because the margin (taken as a luckiness function) would come out typically larger. Nevertheless, the *quantitative* agreement with the true learning curves is usually less good. Hence, an application of the bounds to model selection may in some cases lead to suboptimal results.

On the other hand, the power of the Statistical Mechanics approach comes from the fact that (in the so far limited situations, where it can be applied) it yields *quantitatively exact* results in the thermodynamic limit, with excellent agreement with the simulations of large systems. Hence, this approach can be used to check the tightness of bounds in controlled analytical experiments. We hope that it will also give an idea how bounds could be improved or replaced by good heuristics.

So far, we have restricted our results to a noise free scenario, but it is straightforward to extend the approach to noisy data. It is also possible to include SVM training with errors (resulting in the more advanced optimization problem with slack variables) in the formalism. We expect that our analysis will give insight into the performance of model selection criteria which are used in order to tune the parameters of the SVM learning algorithm to the noise. We have already shown for the noise free case that a very simple statistics like the relative number of support

vectors can give a wrong prediction for the rate of convergence of the generalization error. It will be interesting to see if more sophisticated estimates based on the margin will give tighter bounds.

Acknowledgements

This work was supported by a grant (Op 45/5-2) of the Deutsche Forschungsgemeinschaft and by the British-German Academic Research Collaboration Programme project 1037 of the British council. The work of HS was supported in part by the USA-Israel Binational Science Foundation.

References

1. A. Buhot and M. B. Gordon. Statistical mechanics of support vector machines. In *ESANN'99 – European Symposium on Artificial Neural Networks Proceedings*, pages 201–206. Michel Verleysen, 1999.
2. R. Dietrich, M. Opper, and H. Sompolinsky. Statistical mechanics of support vector networks. *Physical Review Letters*, 82(14):2975–2978, 1999.
3. M. Opper et al. On the annealed VC entropy for margin classifiers: A statistical mechanics study. In B. Schölkopf et al., editor, *Advances in kernel methods – SV Machines*. MIT Press, Cambridge MA, 1999.
4. E. Gardner. The space of interactions in neural networks. *Journal of Physics A*, 21:257–70, 1988.
5. R. Kühn and J.L. van Hemmen. Collective phenomena in neural networks. In E. Domany J.L. van Hemmen and K. Schulten, editors, *Physics of Neural Networks I*. Springer Verlag, New York, 1996.
6. M. Mézard, G. Parisi, and M.G. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.
7. M. Opper and W. Kinzel. Physics of generalization. In E. Domany J. L. van Hemmen and K. Schulten, editors, *Physics of Neural Networks III*. Springer Verlag, New York, 1996.
8. B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Generalization bounds via eigenvalues of the Gram matrix. Submitted to COLT99, February 1999.
9. H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992.
10. J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.
11. A. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.
12. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
13. T. L. H. Watkin, A. Rau, and M. Biehl. The statistical mechanics of learning

a rule. *Reviews of Modern Physics*, 65:499–556, 1993.