# Support vector machines learning noisy polynomial rules

M. Opper[1] and R. Urbanczik[2]
[1]Department of Computer Science and Applied Mathematics,
Aston University, Birmingham B4 7ET, United Kingdom
[2]Institut für theoretische Physik,
Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany

**Abstract**

Using statistical physics, we study support vector machines (SVMs) learning noisy target rules in cases when the optimal predictor is a polynomial of the inputs. If the kernel of the SVM has sufficiently high order or is transcendental, the scale of the learning curve and the asymptote is determined by the target rule and does not depend on the kernel. On this scale we find convergence to optimal generalization but no convergence of the training error to the generalization error.

In recent years support vector machines (SVMs) have become an important alternative to multilayer neural networks for supervised learning from random examples [1, 2]. Their main advantage is that the learning task is mapped onto a convex optimization problem which can be reliably solved even when the function implemented by the SVM is complicated. The basic idea of SVMs is to nonlinearly transform an input $y$ into a feature vector $\Psi(y)$ which is an element of a Hilbert space, and an SVM simply defines an oriented hyperplane $\mathcal{P}$ in this space. In the application to binary classification tasks, the input $y$ is then classified by asking on which side of the hyperplane $\mathcal{P}$ the image $\Psi(y)$ of $y$ lies. This is analogous to the well known approach in nonlinear regression to use a model which is linear in the parameters but nonlinear in the inputs. The distinguishing aspects of support vector machines, however, arise from the way the hyperplane is chosen, based on a training set of $m$ inputs $x^\mu$ and their desired classifications $\tau^\mu$. SVMs construct the hyperplane $\mathcal{P}$ which classifies the training data correctly and which has maximal distance to the images $\Psi(x^\mu)$ of the points in the training set. This maximal distance is called the maximal margin, and it is geometrically intuitive that the maximization tends to improve the odds of classifying a new input correctly. Just as important, it also helps to control the computational complexity of the procedure: The maximal margin hyperplane $\mathcal{P}$ can be expressed as a linear combination of the feature

vectors $\Psi(x^\mu)$, and to classify an input $y$, that is to decide on which side of $\mathcal{P}$ the image $\Psi(y)$ lies, one basically has to evaluate inner products $\Psi(x^\mu) \cdot \Psi(y)$. One now carefully chooses the mapping $\Psi$ and the Hilbert space so that inner products $\Psi(x) \cdot \Psi(y)$ can be evaluated efficiently using a kernel function $k(x, y) = \Psi(x) \cdot \Psi(y)$, without having to individually calculate the feature vectors $\Psi(x)$ and $\Psi(y)$. In this manner it becomes computationally feasible to use very high and even infinite dimensional feature vectors.

This raises the intriguing question whether the use of a very high dimensional feature space may typically be helpful, and recent results [3, 4] obtained by a Statistical Mechanics analysis have been largely negative in this respect. They suggest that it is rather important to match the complexity of the kernel to the target rule. The analysis in [3] considers the case of $N$-dimensional inputs with binary components and assumes that the target rule giving the correct classification $\tau$ of an input $x$ is obtained as the sign of a function $t(x)$ which is polynomial in the input components and of degree $L$. The SVM uses a kernel which is a polynomial of the inner product $x \cdot y$ in input space of degree $K \geq L$, and the feature space dimension is thus $\mathcal{O}(N^K)$. In this scenario it is shown, under mild regularity condition on the kernel and for large $N$, that the SVM generalizes well when the number of training examples $m$ is on the order of $N^L$. So the scale of the learning curve is determined by the complexity of the target rule and not by the kernel. However, considering the rate with which the generalization error approaches zero one finds the optimal $N^L/m$ decay only when $K$ is equal to $L$ and the convergence is substantially slower when $K > L$. So it is important to match the complexity of the kernel to the target rule and using a large value of $K$ is only justified if $L$ is assumed large and if one can use on the order of $N^L$ examples for training.

Here we show that the situation is very different when one considers the arguably more realistic scenario of a target rule corrupted by noise. Now one can no longer use $K = L$ since no separating hyperplane $\mathcal{P}$ will exist when $m$ is sufficiently large compared to $N^L$. However when $K > L$ the SVM exists and we show that it achieves optimal generalization performance in the limit that $N^L/m$ is small. Remarkably, the asymptotic rate of decay of the generalization error is independent of the kernel in this case and a general characterization of the asymptote in terms of properties of the target rule is possible. In a second step we show that under mild regularity conditions these findings also hold when $k(x, y)$ is an arbitrary function of $x \cdot y$ or when the kernel is a function of the Euclidean distance $||x - y||$. The latter type of kernels is widely used in practical applications of SVMs.

While these main results have already been published in [5], the scope of this contribution is to give a more detailed account of their derivation.

We begin by assuming a polynomial kernel $k(x,y) = f(x \cdot y)$ where $f(z) = \sum_{k=0}^{K} \mu_k z^k$ is of degree $K$. Denoting by $\rho$ a multi-index $\rho = (\rho_1, \ldots, \rho_N)$ with $\rho_i \in \mathbb{N}_0$, we set $x_\rho = \sqrt{n_\rho} \prod_{i=1}^{N} x_i^{\rho_i}$ where $n_\rho = |\rho|! / \prod_{i=1}^{N} \rho_i!$ and the degree of $x_\rho$ is $|\rho| = \sum_{i=1}^{N} \rho_i$. The kernel can then be described by features $\Psi_\rho(x) = \sqrt{\mu_{|\rho|}} x_\rho$ since $k(x,y) = \sum_\rho \Psi_\rho(x) \Psi_\rho(y)$, where the summation runs over all multi-indices of degree up to $K$. To assure that the features are real, we assume that the coefficients $\mu_k$ in the kernel are nonnegative. A hyperplane in feature space is parameterized by a weight vector $w$ with components $w_\rho$, and if $0 < \tau^\mu w \cdot \Psi(x^\mu)$, a point $(x^\mu, \tau^\mu)$ of the training set lies on the correct side of the plane. To express that the plane $\mathcal{P}$ has maximal distance to the points of the training set, we choose an arbitrary positive stability parameter $\kappa$ and require that the weight vector $w^*$ of $\mathcal{P}$ minimize $w \cdot w$ subject to the constraints $\kappa < \tau^\mu w \cdot \Psi(x^\mu)$, for $\mu = 1, \ldots, m$. Statistical Mechanics is applied by first analyzing a soft version of the optimization problem characterized by an inverse temperature $\beta$. One considers the partition function

$$Z = \int \mathrm{d}w \, e^{-\frac{1}{2}\beta w \cdot w} \prod_{\mu=1}^{m} \Theta(\tau^\mu w \cdot \Psi(x^\mu) - \kappa), \tag{1}$$

where the constraints are enforced strictly using the Heaviside step function $\Theta$. The properties of $w^*$ are then obtained by evaluating $\ln Z$ and taking the limit $\beta \to \infty$.

To model the training data, we assume that the random and independent input components have zero mean and variance $1/N$. This scaling assures that the variance of $w \cdot \Psi(x^\mu)$ stays finite in the large $N$ limit. For the target rule we assume that its deterministic part is given by the polynomial $t(x) = \sum_\rho \sqrt{\mu_{|\rho|}} B_\rho x_\rho$ with real parameters $B_\rho$. The magnitude of the contribution of each degree $k$ to the value of $t(x)$ is measured by the quantities

$$T_k = \mu_k \frac{1}{N_k} \sum_{\rho, |\rho| = k} B_\rho^2 \tag{2}$$

where $N_k = \binom{N+k-1}{k}$ is the number of terms in the sum. The degree of $t(x)$ is $L$ and lower than $K$, so $T_L > 0$ and $T_{L+1} = \ldots = T_K = 0$. Note, that this definition of $t(x)$ ensures that any feature necessary for computing $t(x)$ is available to the SVM. For brevity we assume that the constant term in $t(x)$ vanishes ($T_0 = 0$) and the normalization is $\sum_k T_k = 1$.

In the deterministic case the label of a point $x$ would simply be the sign of $t(x)$.

Here we consider a nondeterministic rule and the output label is obtained using a random variable $\tau_u \in \{-1,1\}$ parameterized by a scalar $u$. The observable instances of the rule, and in particular the elements of the training set, are then obtained by independently sampling the random variable $(x, \tau_{t(x)})$. Simple examples are additive noise, $\tau_{t(x)} = \mathrm{sgn}(t(x)+\eta)$, or multiplicative noise, $\tau_{t(x)} = \mathrm{sgn}(t(x)\eta)$, where $\eta$ is a noise term independent of $x$. In general, we will assume that the noise does not systematically corrupt the deterministic component, formally

$$\mathrm{Prob}\left(\tau_u = \mathrm{sgn}(u)\right) > \frac{1}{2} \quad \text{for all } u. \tag{3}$$

So $\mathrm{sgn}(t(x))$ is the best possible prediction of the output label of $x$. Further we assume that that the rule is indeed noisy, so the minimal achievable generalization error $\epsilon_{\min}$ is positive:

$$\epsilon_{\min} = \langle \Theta(-t(x)\tau_{t(x)}) \rangle_x > 0. \tag{4}$$

In the limit of many input dimensions $N$, a central limit argument yields that for a typical target rule $\epsilon_{\min} = 2\langle \Theta(-u)\hat{\Theta}(u)\rangle_u$, where u is zero mean and unit variance Gaussian. The function $\hat{\Theta}$ will play a considerable rôle in the sequel. It is a symmetrized form of the probability $p(u)$ that $\tau_u$ is equal to 1, $\hat{\Theta}(u) = \frac{1}{2}(p(u) + 1 - p(-u))$.

One now evaluates the quenched average of $\ln Z$ (Eq. 1) in terms of the replica-symmetric order parameters

$$\begin{aligned} R_k &= \frac{\mu_l}{N_l} \sum_{|\rho|=k} \langle w_\rho \rangle_w B_\rho \\ Q_k &= \frac{\mu_l}{N_l} \sum_{|\rho|=k} \left\langle (w_\rho)^2 \right\rangle_w \\ q_k &= \frac{\mu_l}{N_l} \sum_{|\rho|=k} \langle w_\rho \rangle_w^2 \,, \end{aligned} \tag{5}$$

where the thermal average over $w$ refers to the Gibbs distribution (1). In the limit that both $m$ and $N$ are large, one then finds for the quenched average of $\ln Z$:

$$\begin{aligned} \langle\langle \ln Z \rangle\rangle &\simeq \operatorname*{extr}_{R_k,Q_k,q_k} \; m G^{\mathrm{r}}(R,Q,q) + \sum_{k=1}^{K} N_k G_k^{\mathrm{s}}(R_k,Q_k,q_k) \,, \\ G^{\mathrm{r}}(R,Q,q) &= 2 \left\langle \hat{\Theta}(v) \ln H\left( -\frac{Rv + \sqrt{q-R^2}u - \kappa}{\sqrt{Q-q}} \right) \right\rangle_{u,v} . \end{aligned} \tag{6}$$

Here $u, v$ are independent Gaussian random variables with zero mean and unit variance, and

$$R \equiv \sum_{k=1}^{K} R_k \,, \quad Q \equiv \sum_{k=1}^{K} Q_k \,, \quad q \equiv \sum_{k=1}^{K} q_k \,. \tag{7}$$

The entropy terms $G_k^{\rm s}$ are very similar to what is found for the simple perceptron:

$$2G_k^{\rm s}(R_k,Q_k,q_k) = -\beta Q_k/\mu_k + \frac{q_k - R_k^2/T_k}{Q_k - q_k} + \ln(Q_k - q_k)\,. \tag{8}$$

The solution of (6) will depend on the relative magnitude of $m$ to $N$ and we make the generic Ansatz $m = \alpha N_l$, where $l = 1,\ldots,L$. For $k \neq l$ the values of the $k$-th degree order parameters are then obtained by scaling arguments. For $k > l$ we obtain that to leading order in $N$:

$$R_k = q_k = 0 \quad \text{and } Q_k = \mu_k/\beta\,. \tag{9}$$

The case $k < l$ is more involved and leads to

$$R_k = T_k s, \quad \text{and } Q_k = q_k = T_k s^2 \tag{10}$$

where $s$ is given by the stationarity condition: $\left(2s(\frac{\partial}{\partial Q} + \frac{\partial}{\partial q}) + \frac{\partial}{\partial R}\right) G^{\rm r} = 0$.

Substituting these relations in (6) we obtain that in the large $N$ limit

$$\begin{aligned} \frac{\langle\langle \ln Z\rangle\rangle}{N_l} &= \operatorname*{extr}_{R,Q,q} \alpha G^{\rm r}(R,q,Q) + G_l^{\rm r}(R - t_l s, q - t_l s^2, Q - t_l s^2 - S_l/\beta) \\ s &\equiv \frac{R\mu_l}{S_l T_l + (t_l + T_l)\mu_l + T_l\beta(q-Q)}\,, \end{aligned} \tag{11}$$

where $t_l = \sum_{k=1}^{l-1} T_k$ and

$$S_l = f(1) - \sum_{i=0}^{l} \mu_i\,. \tag{12}$$

To focus on the limit of large $\beta$, where the density on the weight vectors converges to a delta peak and $q \to Q$, we introduce the rescaled order parameter $\chi = \beta(Q-q)/S_l$. Note that this scaling with $S_l$ is only possible since the degree $K$ of the kernel $f(x\cdot y)$ is greater than $l$, and thus $S_l \neq 0$. In this limit we obtain for the typical weight vector $w^*$ of the SVM

$$\begin{aligned} \frac{S_l}{N_l}\left\langle\left\langle -\frac{1}{2} w^*\cdot w^*\right\rangle\right\rangle = \operatorname*{extr}_{r,\chi,q} \\ \frac{-\alpha q}{\chi}\left\langle \hat\Theta(-u) G\left(ru + \sqrt{1-r^2}v - \frac{\kappa}{\sqrt{q}}\right)\right\rangle_{u,v} - \\ \frac{q}{2}\left(\frac{S_l}{\mu_l} - \frac{1}{\chi - 1}\right)\left(1 - \frac{r^2}{-(\chi-1)T_l S_l/\mu_l + \sum_{i=1}^{l} T_i}\right) \end{aligned} \tag{13}$$

where $G(z) = \Theta(z) z^2$, and the physical interpretation of $r$ is that $r = R/\sqrt{Q}$.

Since the stationary value of (13) is finite, the typical length of $w^*$ is of the order $\sqrt{N_l}$. So the higher order components of $w^*$ are small, $(w_\rho^*)^2 \ll 1$ for $|\rho| > l$, although these

components play a crucial rôle in ensuring that a hyperplane separating the training points exists even for large $\alpha$. But the key quantity obtained from Eq. (13) is the stationary value of $r$ which determines the generalization error of the SVM via $\epsilon_g = \langle\hat{\Theta}(-u)\Theta(ru+\sqrt{1-r^2}v)\rangle_{u,v}$, and in particular $\epsilon_g = \epsilon_{\min}$ for $r = 1$.

We now specialize to the case that $l$ equals $L$, the degree of the polynomial $t(x)$ in the target rule. So $m = \alpha N_L$ and for large $\alpha$, after some algebra, Eq. (6) yields

$$r = 1 - \frac{A(q^*)}{4B(q^*)^2}\frac{1}{\alpha} \tag{14}$$

where $B(q) = \left\langle \hat{\Theta}(u)\Theta\left(-u+\kappa/\sqrt{q}\right)\right\rangle_u$, $A(q) = \left\langle \hat{\Theta}(u)\Theta\left(-u+\kappa/\sqrt{q}\right)\left(-u+\kappa/\sqrt{q}\right)^2\right\rangle_u$ and $q^* = \arg\min_q qA(q)$. It is instructive to see that the conditions on our noise model do indeed guarantee that the minimization problem has a solution. We denote the derivative of $qA(q)$ w.r.t. $q$ by $z(q)$, find $z(q) = \left\langle \hat{\Theta}(u)\Theta\left(-u+\kappa/\sqrt{q}\right)\left(u-\kappa/\sqrt{q}\right)u\right\rangle_u$, and show that $z(q)$ is positive for large, and negative for small $q$. In the limit $q \to \infty$ we get $z(\infty) = \left\langle \hat{\Theta}(u)\Theta(-u)u^2\right\rangle_u$, this cannot be negative and is in fact strictly larger than zero since we have assumed that $\epsilon_{\min} > 0$ (Condition 4). On the other hand $z(q)$ is negative for small $q$, since $\sqrt{q}z(q)$ for $q \to 0$ converges to $-\kappa\left\langle \hat{\Theta}(u)u\right\rangle_u$. Condition (3) is easily seen to imply that $\hat{\Theta}(u) > \hat{\Theta}(-u)$ if $u > 0$, and thus $\left\langle \hat{\Theta}(u)u\right\rangle_u$ is positive.

Equation (14) shows that optimal generalization performance is achieved on scale given by the complexity of the target rule in the limit of large $\alpha$. Remarkably, as long as $K > L$, the asymptote is invariant to the choice of the kernel since $A(q)$ and $B(q)$ are defined solely in terms of the target rule.

Our next goal is to understand cases where the kernel is a general function of the inner product or of the distance between the vectors. The kernel must be positive semidefinite so that Mercer's theorem assures that $k(x,y) = \Psi(x)\cdot\Psi(y)$ for a suitable mapping of the inputs into a Hilbert space. We still assume that the target rule is of finite complexity, i.e. defined by a polynomial and corrupted by noise and that the number of examples is polynomial in $N$. Remarkably, the more general kernels then reduce to the polynomial case in the thermodynamic limit.

Since it is difficult to find a description of the Hilbert space for $k(x,y)$ which is useful for a statistical mechanics calculation, our starting point is the dual representation: The weight vector $w^*$ defining the maximal margin hyperplane $\mathcal{P}$ can be written as a linear combination of the feature vectors $\Psi(x^\mu)$ and hence $w^*\cdot\Psi(y) = \sigma(y)$, where

$$\sigma(y) = \sum_{\mu=1}^{m}\lambda_\mu\tau^\mu k(x^\mu, y). \tag{15}$$

By standard results of convex optimization theory the $\lambda_\mu$ are uniquely defined by the Kuhn-Tucker conditions $\lambda_\mu \geq 0$, $\tau^\mu \sigma(x^\mu) \geq \kappa$ (for all patterns), further requiring that for positive $\lambda_\mu$ the second of the two inequalities holds as an equality. One also finds that $w^* \cdot w^* = \sum_{\mu=1}^m \lambda_\mu$ and for a polynomial kernel we thus obtain a bound on $\sum_{\mu=1}^m \lambda_\mu$ since $w^* \cdot w^*$ is $\mathcal{O}(m)$.

We first consider kernels $\phi(x \cdot y)$, with a general continuous function $\phi$ of the inner product, and assume that $\phi$ can be approximated by a polynomial $f$ in the sense that $\phi(1) = f(1)$ and $\phi(z) - f(z) = \mathcal{O}(z^K)$ for $z \to 0$. Now, with a probability approaching 1 with increasing $N$, the magnitude of $x^\mu \cdot x^\nu$ is smaller than, say, $N^{-1/3}$ for all different indices $\mu$ and $\nu$ as long as $m$ is polynomial in $N$. So, considering Eq. (15), for large $N$ the functions $\phi(z)$ and $f(z)$ will only be evaluated in a small region around zero and at $z = 1$ when used as kernels of a SVM trained on $m = \alpha N_L$ examples. Thus for large $N$ and $K > 3L$ the solution of the Kuhn-Tucker conditions for $f$ converges to the one for $\phi$. So Eqs. (13,14) can be used to calculate the generalization error for $\phi$ by setting $\mu_l = \phi^{(l)}(0)/l!$ for $l = 1, \ldots, L$, when $\phi$ is analytical . Note that results in [3] assure that $\mu_l \geq 0$ if the kernel $\phi(x \cdot y)$ is positive definite for all input dimensions $N$. Further, the same reduction to the polynomial case will hold in many instances where $\phi$ is not analytical but just sufficiently smooth close to 0.

We next turn to radial basis function (RBF) kernels where $k(x, y)$ depends only on the Euclidean distance between two inputs, $k(x, y) = \Phi(|x - y|^2)$. For binary input components ($x_i = \pm N^{-1/2}$) this is just the inner product case since $\Phi(|x - y|^2) = \Phi(2 - 2x \cdot y)$. However, for more general input distributions, e.g. Gaussian input components, the fluctuations of $|x|$ around its mean value 1 have the same magnitude as $x \cdot y$ even for large $N$, and an equivalence with inner product kernels is not evident.

Our starting point is the observation that any kernel $\Phi(|x - y|^2)$ which is positive definite for all input dimensions $N$ is a positive mixture of Gaussians [6]. More precisely $\Phi(z) = \int_0^\infty e^{-k^2 z} \, \mathrm{d}a(k)$ where the transform $a(k)$ is nondecreasing. For the special case of a single Gaussian one easily obtains features $\Psi_\rho$ by rewriting $\Phi(|x - y|^2) = e^{-|x-y|^2/2} = e^{-|x|^2/2} e^{x \cdot y} e^{-|y|^2/2}$. Expanding the kernel $e^{x \cdot y}$ into polynomial features, yields the features $\Psi_\rho(x) = e^{-|x|^2/2} x_\rho / \sqrt{|\rho|!}$ for $\Phi(|x - y|^2)$. But, for a generic weight vector $w$ in feature space,

$$w \cdot \Psi(x) = \sum_\rho w_\rho \Psi_\rho(x) = e^{-\frac{1}{2}|x|^2} \sum_\rho w_\rho \frac{x_\rho}{\sqrt{|\rho|!}} \tag{16}$$

is of order 1, and thus for large $N$ the fluctuations of $|x|$ can be neglected.

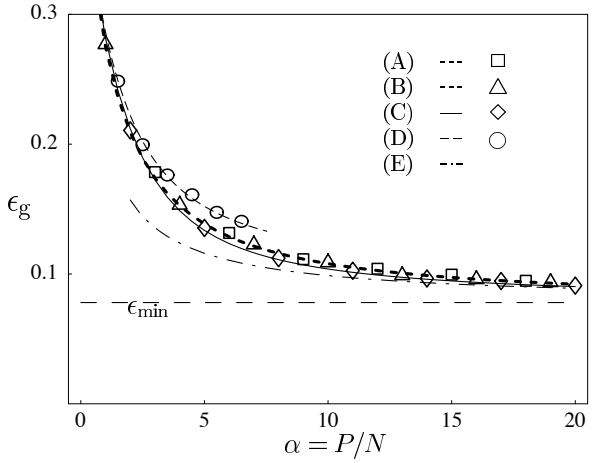


Figure 1: Linear target rule corrupted by additive Gaussian noise $\eta$ ($\langle\eta\rangle = 0$, $\langle\eta^2\rangle = 1/16$) and learned using different kernels. The curves are the theoretical prediction; symbols show simulation results for $N = 600$ with Gaussian inputs and error bars are approximately the size of the symbols. **(A)** Gaussian kernel, $\Phi(z) = e^{-kz}$ with k = 2/3. **(B)** Wiener kernel given by the nonanalytic function $\Phi(z) = e^{-c\sqrt{z}}$. We chose $c \approx 0.065$ so that the theoretical prediction for this case coincides with (A). **(C)** Gaussian kernel with $k = 1/20$, the influence of the parameter change on the learning curve is minimal. **(D)** Perceptron, $\phi(z) = z$. Above $\alpha_c \approx 7.5$ vanishing training error cannot be achieved and the SVM is undefined. **(E)** Kernel invariant asymptote for (A,B,C).

This line of argument can be extended to the case that the kernel is a finite mixture of Gaussians, $\Phi(z) = \sum_{i=1}^{n} a_i e^{-\gamma_i^2 z/2}$ with positive coefficients $a_i$. Applying the reasoning for a single Gaussian to each term in the sum, one obtains a doubly indexed feature vector with components $\Psi_{\rho,i}(x) = e^{-\gamma_i^2|x|^2/2}(a_i\gamma_i^{2|\rho|}/|\rho|!)^{1/2}x_\rho$. It is then straightforward to adapt the calculation of the partition function (Eq. 1 - 13) to the doubly indexed features, showing that the kernel $\Phi(|x-y|^2)$ yields the same generalization behavior as the inner product kernel $\Phi(2-2x\cdot y)$. Based on the calculation, we expect the same equivalence to hold for general radial basis function kernels, i.e. an infinite mixture of Gaussians, even if it would be involved to prove that the limit of many Gaussians commutes with the large $N$ limit.

To illustrate the general results we first consider a scenario where a linear target rule, corrupted by additive Gaussian noise, is learned using different transcendental RBF kernels (Fig. 1). While Eq. (14) predicts that the asymptote of the generalization error does not depend on the kernel, remarkably, the dependence on the kernel is very weak for all values of $\alpha$. In contrast, the generalization error depends substantially on the nature of the noise process. Figure 2 shows the finding for a quadratic rule with

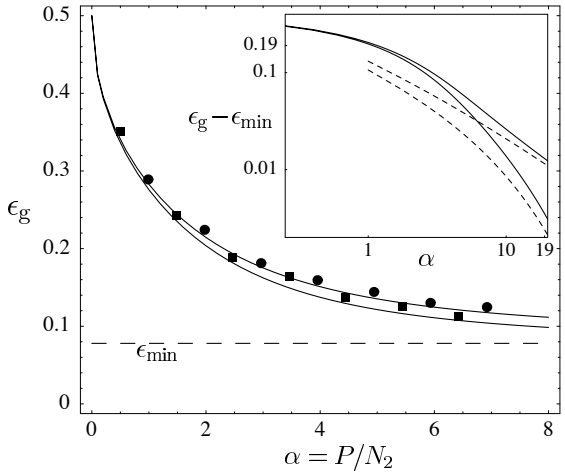


Figure 2: A noisy quadratic rule ($T_1 = 0, T_2 = 1$) learned using the Gaussian kernel with $k = 1/20$. The upper curve (simulations ●) is for additive Gaussian noise as in Fig. 1. The lower curve (simulations ■) is for binary noise, $\eta = \pm s$ with equal probability. We chose $s \approx 0.20$ so that the value of $\epsilon_{\min}$ is the same for the two noise processes. The inset shows that $\epsilon_g$ decays as $1/\alpha$ for Gaussian noise, whereas an exponential decay is found in the binary case. The dashed curves are the kernel invariant asymptotes. The simulations are for $N = 90$ with Gaussian inputs, and standard errors are approximately the size of the symbols.

additive noise for the cases that the noise is Gaussian and binary. In the Gaussian case a $1/\alpha$ decay of $\epsilon_g$ to $\epsilon_{\min}$ is found, whereas for binary noise the decay is exponential in $\alpha$. Note that in both cases the order parameter $r$ approaches 1 as $1/\alpha$.

The general characterization of learning curves obtained in this Paper demonstrates that support vector machines with high order or even transcendental kernels have definitive advantages when the training data is noisy. Further Eq. (9) shows that maximizing the margin is an essential ingredient of the approach: If one just chooses any hyperplane which classifies the training data correctly, the scale of the learning curve is not determined by the target rule and far more examples are needed to achieve good generalization. Nevertheless our findings are at odds with many of the current theoretical motivations for maximizing the margin which argue that this minimizes the effective Vapnik Chervonenkis dimension of the classifier and thus ensures fast convergence of the training error to the generalization error [1, 2]. Since we have considered hard margins, in contrast to the generalization error, the training error is zero, and we find no convergence between the two quantities. But close to optimal generalization is achieved since maximizing the margin biases the SVM to explain as much as possible of the data in terms of a low order polynomial. While the Statistical Physics approach used is only

exactly valid in the thermodynamic limit, the numerical simulations indicate that the theory is already a good approximation for a realistic number of input dimensions.

We thank Rainer Dietrich for useful discussion and for giving us his code for the simulations. The work of M.O. was supported by the EPSRC (grant no. GR/M81608) and the British Council (ARC project 1037); R.U. was supported by the DFG and the DAAD.

## References

[1] V. Vapnik. *The nature of statistical learning theory.* Springer, New York, 1995.

[2] N. Cristianini and J. Shawe-Taylor. *Support Vector Machines.* Cambridge University Press, 2000.

[3] R. Dietrich, M. Opper, and H. Sompolinsky. Statistical mechanics of support vector networks. *Phys. Rev. Lett.*, 82:2975 –2978, 1999.

[4] S. Risau-Gusman and M. Gordon. Generalization properties of finite-size polynomial support vector machines. *Phys. Rev. E*, 62:7092–7099, 2000.

[5] M. Opper and R. Urbanczik. Universal learning curves of support vector machines. *Phys. Rev. Lett.*, 86:4410–4413, 2001.

[6] I. Schoenberg. Metric spaces and completely monotone functions. *Anal. Math*, 39:811–841, 1938.