

ANNEALED THEORIES OF LEARNING

H. S. Seung
AT&T Bell Laboratories
Murray Hill, NJ 07974
E-mail: `seung@physics.att.com`

ABSTRACT

We study annealed theories of learning boolean functions using a concept class of finite cardinality. The naive annealed theory can be used to derive a universal learning curve bound for zero temperature learning, similar to the inverse square root bound from the Vapnik-Chervonenkis theory. Tighter, nonuniversal learning curve bounds are also derived. A more refined annealed theory leads to still tighter bounds, which in some cases are very similar to results previously obtained using one-step replica symmetry breaking.

1. Introduction

The annealed approximation¹ has proven to be an invaluable tool for studying the statistical mechanics of learning from examples. Previously it was found that the annealed approximation gave qualitatively correct results for several models of perceptrons learning realizable rules.² Because of its simplicity relative to the full quenched theory, the annealed approximation has since been used in studies of more complicated multilayer architectures.^{3,4} However, it was also found that the application of the annealed approximation to learning of unrealizable rules could yield very wrong behavior.² To obtain correct learning curve asymptotics, the more complex quenched theory was needed.

In this paper, we revisit the subject of annealed theories of learning, in the special case where the concept class consists of boolean-valued functions with discrete weights. Using annealed arguments, we construct upper bounds on the generalization error that compare favorably to the quenched theory results. The naive annealed theory can be used to derive a universal learning curve bound of an inverse square root law, similar to the bound from the Vapnik-Chervonenkis theory.⁵ It can also be used to derive tighter, nonuniversal learning curve bounds.

Even tighter bounds can be obtained by using a different annealed theory. We study one that is based on the difference in the training errors of a concept and the optimal concept in the class. This annealed theory can in some cases produce learning curve bounds that are comparable to those of replica symmetry breaking calculations. This paper is in part a translation into physicists' language of results

that have previously appeared elsewhere in more mathematical form.⁶

2. Notational preliminaries

Let \mathcal{W} be a class of boolean-valued functions w , or concepts, with finite cardinality $|\mathcal{W}| = 2^N$. There is a target concept, or teacher, that is to be learned. The error function relative to the target concept on an input x is defined by

$$e(w, x) = \begin{cases} 1, & \text{if } w \text{ disagrees with the target on } x \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The training error of a concept w on a sample of inputs $X^m = (x_1, \dots, x_m)$ is defined as the frequency of errors on the sample,

$$e_t(w, X^m) = \frac{1}{m} \sum_{i=1}^m e(w, x_i) . \quad (2)$$

The m points in the sample are assumed to be drawn independently at random from an input distribution. The generalization error of a concept w is defined as the probability that it disagrees with the target concept on an input drawn at random from this distribution. Equivalently, we may write it as the expectation of $e(w, x)$ with respect to the input distribution, denoted by double angle brackets

$$e_g(w) = \langle\langle e(w, x) \rangle\rangle . \quad (3)$$

We assume that $e_g(w)$ possesses a unique minimum w^* in the concept class with generalization error $\epsilon^* = e_g(w^*)$. If $\epsilon^* = 0$ then the target concept is said to be realizable by the concept class. Otherwise the target concept is unrealizable.

3. Quenched theory

Suppose the learner has chosen a hypothesis with training error ϵ_t on a given sample X^m . What can we say about its generalization error? The learner's hypothesis is just one member of the set of all hypotheses with training error ϵ_t . If we were to go through this whole set of hypotheses and tabulate their generalization errors, we would obtain a histogram. As a function of ϵ_g , the entropy $s(X^m, \epsilon_t, \epsilon_g)$ is this histogram on a log scale^a

$$s(X^m, \epsilon_t, \epsilon_g) = \frac{1}{N} \log \sum_w \delta(\epsilon_g - e_g(w)) \delta(\epsilon_t - e_t(w, X^m)) . \quad (4)$$

^aAlthough some bin width must be chosen for the histogram, the entropy is independent of this width in the thermodynamic limit, so we simply write delta functions in its definition. There are also some technical subtleties due to the fact that the argument of the logarithm can vanish, difficulties that are finessed by the replica formalism.

In the thermodynamic limit $m, N \rightarrow \infty$ with $\alpha \equiv m/N$ constant, if the entropy becomes self-averaging it can be replaced by its expectation over the random sample X^m ,

$$s(\alpha, \epsilon_t, \epsilon_g) = \langle\langle s(X^m, \epsilon_t, \epsilon_g) \rangle\rangle . \quad (5)$$

We assume that the entropy becomes a smooth function in the thermodynamic limit, i.e., that the histogram looks smooth on a log scale. However, on a linear scale the histogram is very sharply peaked about its maximum

$$\operatorname{argmax}_{\epsilon_g} s(\alpha, \epsilon_t, \epsilon_g) . \quad (6)$$

This is the typical value of ϵ_g in this set, the generalization error if the learner chooses at random from the set of all hypotheses with training error ϵ_t .

If the learner does not choose completely at random from this set, the generalization error might be different. The generalization errors that are possible are in the region $\{\epsilon_g : s(\alpha, \epsilon_t, \epsilon_g) \geq 0\}$, since nonnegative entropy means that there is at least one hypothesis with this ϵ_t and ϵ_g . Outside this region, there are no hypotheses, and the entropy is $-\infty$. Hence the worst possible generalization is given by

$$\max\{\epsilon_g : s(\alpha, \epsilon_t, \epsilon_g) \geq 0\} . \quad (7)$$

Assuming the entropy approaches zero smoothly, we may simply solve

$$s(\alpha, \epsilon_t, \epsilon_g) = 0 \quad (8)$$

with respect to ϵ_g , and pick the appropriate root. Obviously, the typical generalization error is upper bounded by the worst. Note that if the concept class has infinite cardinality, the zero entropy point does not identify the worst generalization error.

It is convenient to calculate the free energy

$$-\beta f(\alpha, \beta, \epsilon_g) = \frac{1}{N} \langle\langle \log \sum_w \delta(\epsilon_g - e_g(w)) e^{-\beta m \epsilon_t(w, X^m)} \rangle\rangle , \quad (9)$$

and then obtain the entropy via Legendre transformation

$$s(\alpha, \epsilon_t, \epsilon_g) = \min_{\beta} \{\alpha \beta \epsilon_t - \beta f(\alpha, \beta, \epsilon_g)\} , \quad (10)$$

$$\beta f(\alpha, \beta, \epsilon_g) = \min_{\epsilon_t} \{\alpha \beta \epsilon_t - s(\alpha, \epsilon_t, \epsilon_g)\} . \quad (11)$$

When s and βf are smooth, the minima are found by differentiating, so that the relationship between temperature $T = 1/\beta$ and training error ϵ_t is given by

$$\alpha \beta = \left. \frac{\partial s}{\partial \epsilon_t} \right|_{\alpha, \epsilon_g} , \quad (12)$$

$$\alpha \epsilon_t = \left. \frac{\partial(\beta f)}{\partial \beta} \right|_{\alpha, \epsilon_g} . \quad (13)$$

These results imply that the training error is an increasing function of temperature, and that entropy is an increasing function of ϵ_t or T . As the entropy increases, the set of generalization errors with nonnegative entropy in (7) is enlarged, and hence the worst generalization error is an increasing function of training error or temperature^b

Let us review the typical learning curve behavior for unrealizable rules. The zero temperature training error $\epsilon_t(\alpha, T = 0)$ is the lowest possible training error that can be achieved. For small α , all training examples can be loaded perfectly, so that $\epsilon_t(\alpha, T = 0) = 0$. However, for large α $\epsilon_t(\alpha, T = 0) \rightarrow \epsilon^*$ from below, since the rule is unrealizable. For finite temperatures, or training errors above $\epsilon_t(\alpha, T = 0)$, the entropy is positive in some region of ϵ_g . As the training error decreases, the entropy drops, until at $\epsilon_t(\alpha, T = 0)$, there is no region of positive entropy. In other words, $\epsilon_t(\alpha, T = 0)$ is defined as the training error for which the maximum of the entropy is zero,

$$\max_{\epsilon_g} \{s(\alpha, \epsilon_t(\alpha, T = 0), \epsilon_g)\} = 0 . \quad (14)$$

4. Naive annealed theory

By moving the sample average inside the logarithm, we can upper bound the quenched entropy (5) by the annealed entropy,

$$s^{ann}(\alpha, \epsilon_t, \epsilon_g) = \frac{1}{N} \log \langle \langle \sum_w \delta(\epsilon_g - e_g(w)) \delta(\epsilon_t - e_t(w, X)) \rangle \rangle . \quad (15)$$

Any upper bound on the quenched entropy can be used to obtain an upper bound on the worst generalization error (7), a fact that will be used repeatedly throughout this paper. In particular, $s \leq s^{ann}$ implies that

$$\max_{\epsilon_g} \{\epsilon_g : s^{ann}(\alpha, \epsilon_t, \epsilon_g) \geq 0\} \geq \max_{\epsilon_g} \{\epsilon_g : s(\alpha, \epsilon_t, \epsilon_g) \geq 0\} , \quad (16)$$

so that solving $s^{ann}(\alpha, \epsilon_t, \epsilon_g) = 0$ for ϵ_g gives an upper bound on the worst generalization error. To obtain an approximation for the typical generalization error, we maximize the annealed entropy with respect to ϵ_g , in analogy with (6)

$$\operatorname{argmax}_{\epsilon_g} s^{ann}(\alpha, \epsilon_t, \epsilon_g) . \quad (17)$$

Note that the annealed approximation is not a bound on the generalization error; it is an approximation.

^bIn contrast, it appears that the typical generalization error (6) may exhibit nonmonotonic behavior as a function of ϵ_t , although the existence of this *overtraining* phenomenon has not yet been conclusively demonstrated.

The annealed free energy is easier to compute than the entropy, since the sample average factorizes to yield

$$-\beta f^{ann}(\alpha, \beta, \epsilon_g) = \frac{1}{N} \log \sum_w \delta(\epsilon_g(w) - \epsilon_g) \langle \langle e^{-m\beta\epsilon_t(w, X^m)} \rangle \rangle \quad (18)$$

$$= \rho(\epsilon_g) + \alpha \log[1 + (e^{-\beta} - 1)\epsilon_g], \quad (19)$$

where

$$\rho(\epsilon_g) = \frac{1}{N} \log \sum_w \delta(\epsilon_g - \epsilon_g(w)). \quad (20)$$

The annealed entropy follows from the Legendre transformation (10),

$$s^{ann}(\alpha, \epsilon_t, \epsilon_g) = \rho(\epsilon_g) - \alpha D(\epsilon_t || \epsilon_g), \quad (21)$$

where the function $D(p||q)$ is the relative entropy of the two binary distributions $(p, 1-p)$ and $(q, 1-q)$,

$$D(p||q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}. \quad (22)$$

The relationship between temperature $T = 1/\beta$ and training error ϵ_t in the annealed theory can be found by applying equation (12) to obtain

$$\beta = \log \frac{\epsilon_g}{\epsilon_t} - \log \frac{1-\epsilon_g}{1-\epsilon_t}. \quad (23)$$

There are two simple limits to be taken, $T \rightarrow \infty$ and $T = 0$. In the infinite temperature limit, the training and generalization errors are equal, $\epsilon_t = \epsilon_g$. Furthermore, the annealed theory becomes equivalent to the quenched,

$$s^{ann} = -\beta f^{ann} = \rho(\epsilon_g) = s = -\beta f. \quad (24)$$

At zero temperature, the training error is zero, and

$$s^{ann} = -\beta f^{ann} = \rho(\epsilon_g) - \alpha \log(1 - \epsilon_g). \quad (25)$$

So we see that the structure of the annealed theory is simpler than that of the quenched. The annealed entropy (21) is composed of two terms. The first term $\rho(\epsilon_g)$ is the entropy at infinite temperature. The second term $\alpha D(\epsilon_t || \epsilon_g)$ is linear in α and contains the dependence on temperature or training error. Whereas the first term depends on the concept class, target concept, and input distribution, the second term is universal.

What price do we pay for this gain in theoretical simplicity? Seung, Sompolinsky, and Tishby observed that the annealed theory gives qualitatively correct behavior for realizable rules, but can be very wrong for unrealizable rules.² The deficiency of the

annealed theory is rooted in the fact that training error vanishes at zero temperature. Although this is appropriate for realizable rules, zero training error is not attainable for large α in the case of unrealizable rules. In the annealed theory, both the concept w and the sample X^m are thermal variables. Hence the ground state of the system is the minimum of the training energy with respect to both w and X^m . This means that cooling finds an atypical sample for which zero training error can be achieved. For boolean outputs (the case we consider here), the annealed generalization error always approaches ϵ^* as $\alpha \rightarrow \infty$. However, Seung, Sompolinsky, and Tishby noted that the rate of approach in the annealed theory can be very different from the correct quenched theory.²

Here we use the annealed theory to upper bound the worst generalization error at zero temperature. However, we avoid the trap of setting $T = 0$ in the annealed theory, which gives unphysical results. Recall that any upper bound on the quenched entropy yields an upper bound on the worst generalization error. The following string of inequalities upper bounds the quenched entropy at zero temperature,

$$s(\alpha, \epsilon_t(\alpha, T = 0), \epsilon_g) \leq s^{ann}(\alpha, \epsilon_t(\alpha, T = 0), \epsilon_g) \quad (26)$$

$$\leq s^{ann}(\alpha, \epsilon^*, \epsilon_g) \quad (27)$$

$$= \rho(\epsilon_g) - \alpha D(\epsilon^* || \epsilon_g) \quad (28)$$

$$\leq \rho(\epsilon_g) - 2\alpha(\delta\epsilon_g)^2 \quad (29)$$

$$\leq \log 2 - 2\alpha(\delta\epsilon_g)^2 . \quad (30)$$

The second line is due to the fact that the entropy is an increasing function of ϵ_t . The fourth line comes from the standard bound on the relative entropy $D(p||q) \geq 2(p-q)^2$. The last inequality is due to the fact that the total cardinality of the concept class is 2^N , and gives the coarse bound of

$$\delta\epsilon_g \leq \mathcal{O}(\alpha^{-1/2}) . \quad (31)$$

This inverse square root behavior has the same origin as the corresponding VC bound, and does not require a thermodynamic limit. Application of the penultimate inequality requires a knowledge of the behavior of $\rho(\epsilon_g)$ at small $\delta\epsilon_g$. We define an exponent y by $\rho(\epsilon_g) \sim \mathcal{O}((\delta\epsilon_g)^y)$, up to logarithmic corrections. The resulting tighter bound is

$$\delta\epsilon_g \leq \mathcal{O}(\alpha^{-1/(2-y)}) . \quad (32)$$

5. Refined annealed theory

The annealed theory outlined above is just one of an infinite class of possible annealed theories. To see this, note that we can subtract any function of quenched disorder from the energy, and simply shift the free energy by a constant. Let $E(w, x)$

be an energy depending on thermal and quenched variables w and x . The free energy is then given by

$$-\beta F = \langle\langle \log \sum_w e^{-\beta E(w,x)} \rangle\rangle \quad (33)$$

$$= \langle\langle \log e^{-\beta E_0(x)} \sum_w e^{-\beta(E(w,x)-E_0(x))} \rangle\rangle \quad (34)$$

$$= \langle\langle \log \sum_w e^{-\beta(E(w,x)-E_0(x))} \rangle\rangle - \beta \langle\langle E_0(x) \rangle\rangle . \quad (35)$$

This in turn is bounded by the annealed free energy,

$$-\beta F^{\text{ann}} = \log \sum_w \langle\langle e^{-\beta(E(w,x)-E_0(x))} \rangle\rangle - \beta \langle\langle E_0(x) \rangle\rangle . \quad (36)$$

The function $E_0(x)$ must be chosen to make the bound tight and yet easy to calculate.

Especially in the large α limit, we are most interested in concepts w in the vicinity of w^* . The fluctuations in the training errors of these concepts can be correlated with the fluctuations in the training error of w^* . Hence a sensible choice for us is to subtract the training error of w^* ,

$$-\beta f^{\text{ann}}(\alpha, \beta, \epsilon_g) = \frac{1}{N} \log \sum_w \delta(\epsilon_g - e_g(w)) \langle\langle e^{-\beta m e_t(w, X^m) + \beta m e_t(w^*, X^m)} \rangle\rangle - \alpha \beta \epsilon^* . \quad (37)$$

The correlation between the fluctuations of $e_t(w, X^m)$ and $e_t(w^*, X^m)$ depends on the probability of disagreement $e_d(w, w^*)$ between w and w^* . Thus another order parameter ϵ_d appears, so that

$$-\beta f^{\text{ann}}(\alpha, \beta, \epsilon_g, \epsilon_d) = \rho(\epsilon_g, \epsilon_d) + \alpha \log[1 + \epsilon_d(\cosh \beta - 1) - (\delta \epsilon_g) \sinh \beta] - \alpha \beta \epsilon^* , \quad (38)$$

and

$$\rho(\epsilon_g, \epsilon_d) = \frac{1}{N} \log \sum_w \delta(\epsilon_g - e_g(w)) \delta(\epsilon_d - e_d(w, w^*)) . \quad (39)$$

Maximizing these expressions with respect to ϵ_d yields $-\beta f^{\text{ann}}(\alpha, \beta, \epsilon_g)$ and $\rho(\epsilon_g)$.

The annealed entropy is then obtained via Legendre transformation

$$s^{\text{ann}}(\alpha, \epsilon_t, \epsilon_g, \epsilon_d) = \rho(\epsilon_g, \epsilon_d) + \alpha \min_{\beta} \{ \beta \delta \epsilon_t + \log[1 + \epsilon_d(\cosh \beta - 1) - (\delta \epsilon_g) \sinh \beta] \} \quad (40)$$

The relationship (12) between training error and temperature now takes the form

$$\delta \epsilon_t = \frac{\delta \epsilon_g \cosh \beta - \epsilon_d \sinh \beta}{1 + \epsilon_d(\cosh \beta - 1) - \delta \epsilon_g \sinh \beta} . \quad (41)$$

As before, it is useful to investigate the two limits of infinite and zero temperatures. At infinite temperature, $\delta \epsilon_g = \delta \epsilon_t$ and the annealed theory is equivalent to the quenched. At zero temperature, $\delta \epsilon_t = -1$. This is because annealing the examples finds a sample such that $e_t(w^*, X^m) = 1$ and $e_t(w, X^m) = 0$.

As in the naive annealed theory, we can obtain an upper bound for the worst generalization error by using the annealed entropy at $\epsilon_t = \epsilon^*$. This corresponds to a temperature satisfying $\tanh \beta = \delta\epsilon_g/\epsilon_d$,

$$s^{ann}(\alpha, \epsilon_t = \epsilon^*, \epsilon_g, \epsilon_d) = \rho(\epsilon_g, \epsilon_d) + \alpha \log(1 + \sqrt{\epsilon_d^2 - (\delta\epsilon_g)^2} - \epsilon_d) \quad (42)$$

$$\leq \rho(\epsilon_g, \epsilon_d) + \alpha(\delta\epsilon_g)^2/(2\epsilon_d). \quad (43)$$

Now suppose that $e_d(w, w^*) \leq v(e_g(w))$. Then we can write the upper bound

$$s^{ann}(\alpha, \epsilon_t = \epsilon^*, \epsilon_g, \epsilon_d) \leq \rho(\epsilon_g) + \alpha \frac{(\delta\epsilon_g)^2}{2v(\delta\epsilon_g)}. \quad (44)$$

Defining the exponent z by $v(\epsilon_g) \leq \mathcal{O}((\delta\epsilon_g)^z)$, we derive the asymptotic bound

$$\delta\epsilon_g \leq \mathcal{O}(\alpha^{-1/(2-y-z)}), \quad (45)$$

where y is defined as before by $\rho(\delta\epsilon_g) \leq \mathcal{O}((\delta\epsilon_g)^y)$.

6. Specific learning models

Györgyi and Tishby considered a learning model in which the target function was a perceptron whose inputs were corrupted by noise.⁷ The learner was presented with example pairs of an input x drawn from a Gaussian distribution, along with a label given by $\text{sgn}(w^* \cdot (x + \xi))$, where w^* is the weight vector of the target. The noise vector ξ was also assumed to be Gaussian. We consider not Györgyi and Tishby's spherical perceptron, but rather the finite cardinality concept class of Ising perceptrons (with ± 1 weights). In this case it can be shown that $y = 1$ and $z = 1/2$, so that the naive annealed learning curve bound is $\mathcal{O}(\alpha^{-1})$ and the refined annealed bound is $\mathcal{O}(\alpha^{-2})$, up to logarithmic corrections.

For Györgyi and Tishby's spherical perceptron, we cannot give an upper bound, since the zero entropy point means nothing for a concept class of infinite cardinality. However, we can follow the annealed approximation along the $\epsilon_t = \epsilon^*$ line. Since $\rho(\epsilon_g) \sim \log \epsilon_g$, we take $y = 0$ along with $z = 1/2$, yielding a learning curve approximation of $\mathcal{O}(\alpha^{-2/3})$. It would be interesting to see whether the replica symmetry breaking calculation for their model yield similar asymptotics. A similar argument appears to explain the 2/3 power law seen by Kabashima and Shinomoto.⁸

Seung, Sompolinsky, and Tishby studied a perceptron learning model in which the weights of the teacher w^0 are drawn at random from a Gaussian distribution with zero mean and unit variance, but the learner is an Ising perceptron.² For this model, it can be shown that $y = 1/2$ and $z = 1/4$, resulting in the bound $\delta\epsilon_g \leq \mathcal{O}(\alpha^{-4/5})$, which is the same power law as that obtained using one-step replica symmetry breaking.

7. Conclusion

In the refined annealed theory, the large- α asymptotics of learning curve bounds

depend on two exponents y and z defined by $\rho(\epsilon_g) \leq \mathcal{O}((\delta\epsilon_g)^y)$ and $v(\epsilon_g) \leq \mathcal{O}((\delta\epsilon_g)^z)$. These two exponents give rise to a variety of behaviors:

- If $y + z > 2$, there is a first-order (sudden) phase transition .
- If $1 < y + z < 2$, the error decays as a power law, $1/\alpha^{1/(2-y-z)}$.
- In the marginal case $y + z = 2$, the behavior can be affected by logarithmic corrections to the scaling behavior of $\rho(\epsilon_g)$ and $v(\epsilon_g)$. In the absence of such corrections, there is a second-order (continuous) transition. A logarithmic correction results in exponential decay with α .

Previous work on classification of learning curve asymptotics was mostly restricted in applicability to the realizable case.^{1,2,9}

Of course, the above results only concern the behavior of the bounds, which might not resemble the true behavior. However, we have shown that the bound can be very close to the true behavior in some cases, even when the true behavior is affected by replica symmetry breaking.

The most severe limitation of our results is that they apply only to concept classes of finite cardinality. Our results suggest that a better annealed approximation can be constructed for concept classes of infinite cardinality. However, it is not clear how to construct annealed bounds, since the zero entropy point has no meaning. For the infinite cardinality case, replicas have been used to calculate the worst generalization error.¹⁰ Perhaps tight replica-free bounds for the infinite cardinality case will be possible by combining statistical mechanics techniques with tools from the uniform convergence literature like VC entropy and dimension.⁵

8. References

1. D. B. Schwartz, V. K. Samalam, J. S. Denker, and S. A. Solla. Exhaustive learning. *Neural Comput.*, 2:374–385, 1990.
2. H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Phys. Rev.*, A45:6056–6091, 1992.
3. K. Kang, J.-H. Oh, C. Kwon, and Y. Park. Generalization in a two-layer neural network. *Physical Review E*, 48:4805–4809, 1993.
4. H. Schwarze and J. Hertz. Generalization in fully connected committee machines. *Europhys. Lett.*, 21:785–790, 1993.
5. V. N. Vapnik. *Estimation of Dependences based on Empirical Data*. Springer-Verlag, New York, 1982.
6. D. Haussler, M. Kearns, H. S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. In M. K. Warmuth, editor, *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory*, pages 76–87, New York, 1994. ACM.

7. G. Györgyi and N. Tishby. Statistical theory of learning a rule. In W. K. Theumann and R. Köberle, editors, *Neural Networks and Spin Glasses*, pages 3–36, Singapore, 1990. World Scientific.
8. Y. Kabashima and S. Shinomoto. Learning curves for error minimum and maximum likelihood algorithms. *Neural Comput.*, 4:712–719, 1992.
9. S. Amari, N. Fujita, and S. Shinomoto. Four types of learning curves. *Neural Comput.*, 4:605–618, 1992.
10. A. Engel and C. Van den Broeck. Systems that can learn from examples: replica calculation of uniform convergence bounds for the perceptron. *Phys. Rev. Lett.*, 71:1772–1775, 1993.