Estimating Exact Form of Generalisation Errors

Jianfeng Feng

Biomathematics Laboratory, The Babraham Institute, Cambridge CB2 4AT, UK

Abstract. A novel approach to estimate generalisation errors of the simple perceptron of the worst case is introduced. It is well known that the generalisation error of the simple perceptron is of the form d/t with an unknown constant d which depends only on the dimension of inputs, where t is the number of learned examples. Based upon extreme value theory in statistics we obtain an exact form of the generalisation error of the simple perceptron. The method introduced in this paper opens up new possibilities to consider generalisation errors of a class of neural networks.

1 Introduction

Generalisation errors together with learning errors show how fast a learning machine improves its behaviour when the number t of training examples increases. There are several approaches to this problem.

- According to the Vapnik-Chervonenkis(VC) theory of learning curves, minimising empirical error within a function class $\mathcal{F}$ on a random sample of t examples leads to generalisation error bounded by $O(d/t)$ in the case that the target function is contained in $\mathcal{F}$. The bound is universal: it holds for any class of hypothesis function $\mathcal{F}$, for any input distribution and for any target function. The only problem specific quantity remaining in the bound is the VC dimension d, a measure of the complexity of the function class $\mathcal{F}$. There are a lot of research activities along this line, see for example in [4, 16, 24, 25].
- The in-depth statistical mechanical approach is an another origin of research. It proves results for specific models for which tools such as the replica trick can be applied [6, 18, 23, 21], although a rigorous justification for the replica trick has not been provided [20]. Results are true under the thermodynamic limit [26].
- There are statistical and information theoretical methods of approach too [27, 15]. Most of these approaches suggest that the generalisation error decreases universally in the order of $1/t$ with only its coefficient unknown. The group led by Amari ([2, 19]) proposed a rigorous approach to tackle the problem of learning curves: a statistical approach based upon the expansion of estimators and the generalisation error measured by the entropic loss. They proved again that the generalisation error is of the form $1/t$ with an exactly given coefficient of it depending on the dimension m of the input signals.

The theory of generalisation errors is already well developed, however, the exact form of generalisation errors of some concrete learning rule is rarely known[22]. Even in the simplest case–the simple perceptron, the problem to find the coefficient of the generalisation error is still open except for some very special cases[2]. In this letter, based upon the extreme value theory of statistics we propose a novel approach to tackle the problem and open up new possibilities to rigorously consider the generalisation errors of a class of learning machines. The idea underlying our approach is straightforward. The generalisation error for a given machine is universal, as confirmed by all previous studies, in the sense that it does not depend on the input distribution at all. This fact suggests that to calculate the generalisation errors we should consider the input distribution as simple as possible. For a specific input we show that the generalisation error of the simple perceptron is basically a linear combination of extreme values of input signals. Fortunately, for extremes of an i.i.d random sequence[7] we fully understand their properties, which enables us to complete our calculation.

2 Framework

Consider a machine fed with two dimensional independent inputs $\xi(\tau) = (\xi_1(\tau), \xi_2(\tau)) \in \Omega \subset I\!R^2$. Without loss of generality we assume that the task for the machine to accomplish is the classification problem–to separate data set $\{\xi(\tau), \xi_1(\tau) < 0\}$ from $\{\xi(\tau), \xi_1(\tau) > 0\}$ and so $\text{sign}(\xi_1(\tau))$ is the so-called target function(see Remark 3 below). Suppose that after trained by t examples in terms of a learning rule, for example the simple perceptron learning rule, the output of the learned machine is $h(\xi(t+1)) \in \{-1, 1\}$ when a new signal $\xi(t+1)$ is coming. One key assumption of our approach is that we take into account the case of worst learning.

Assumption 1: all new coming signals dropped on the one side of the line $h(x,y), (x,y) \in I\!R^2$ are correctly recognised, whereas on the other side are not correctly classified.

Suppose that the distribution of $\xi(\tau)$ is symmetric with respect to $x = 0$, the generalisation error of two dimensional case can then be defined by

$$\begin{aligned} \epsilon(t,2) &= \langle |h(\xi(t+1)) - \text{sign}(\xi_1(t+1))| \rangle \\ &= \langle P(\xi(t+1) \in \Omega(t) | \mathcal{F}_t) \rangle \end{aligned} \tag{1}$$

where $\Omega(t)$ is the region(the filled region shown in Fig. 1 (b)) between the target function and function h and $\Omega(t) \in \mathcal{F}_t$, $\mathcal{F}_t$ is the sigma-algebra generated by $\xi(\tau), \tau \leq t$.

Let $\xi_1(tk)$ be the $(t-k)$-th smallest minimum in the set $\{\xi_1(\tau), \tau = 1, \cdots, t\}$. and so

$$\begin{aligned} \xi_1(tt) &= \min\{\xi_1(\tau), \tau = 1, \cdots, t\} \\ \xi_1(t(t-1)) &= \min\{\xi_1(\tau) \geq \xi_1(tt), \tau = 1, \cdots, t, \tau \neq tt\} \\ &\vdots \end{aligned} \tag{2}$$

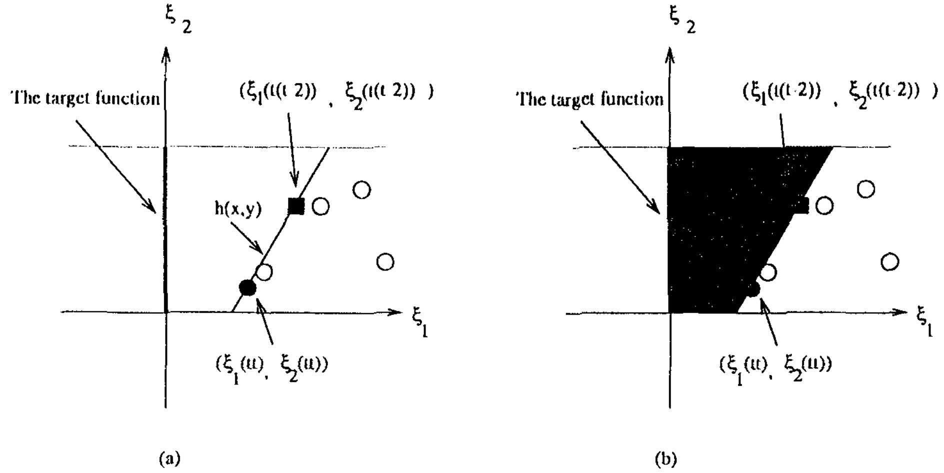
Assumption 1 thus indicates that the line h passes through the minimum $(\xi_1(tt), \xi_2(tt))$ and another k-th smallest minimum. Note that here k depends on the realization of $\{\xi(\tau), \tau = 1, \cdots, t\}$(Fig. 1).

3 Extreme Value Theorey

There are three types of behaviours for extreme values $\xi_1(tt)$ of a sequence of random variables $\xi_1(1), \xi_1(2), \cdots, \xi_1(t)$. For a full exposition of extreme value theory we refer the reader to [17, 13]. Typically for an extreme $\xi_1(t(t-k))$ of a sequence of random variables, i.e. for the k-th minimum of a sequence, we have the following property

$$\langle \xi_1(t(t-k)) \rangle = c(k) o(\gamma(t)) \tag{3}$$

where $c(k)$ is a constant depending on k and $\gamma(t)$ is a vanishing rate of t.



(a)

(b)

Fig. 1. Open circle=examples or input signals. Filled circle=$(\xi_1(tt), \xi_2(tt))$ and filled rectangle=$(\xi_1(t(t-2)), \xi_2(t(t-2)))$. The target function is sign(x). (a). After learning t examples a perceptron is capable of separating data on the two sides of the line $h(x, y)$. (b). Filled region $= \Omega(t)$.

4 Results

Suppose that $\xi_1(\tau) \sim U(0,1)$. When $t \to \infty$ we have

$$\begin{aligned} &i). && P(\xi_1(tt) \geq \frac{x}{t}) = e^{-x} \\ &ii). && P(\xi(t(t-k)) \geq \frac{x}{t}) = e^{-x} \sum_{s=0}^{k-1} \frac{x^s}{s!} \\ &iii). && \langle \xi_1(t(t-k)) \rangle = \frac{k+1}{t} \end{aligned} \tag{4}$$

for $x \geq 0$.

Proof i). From Example 1.7.9 in [17] we know that $P(\eta(tt) \leq 1-x/t) = e^{-x}$ for $\eta(tt)$ representing the largest maximum of $\xi_1(\tau), \tau = 1, \cdots t$. Then i) is a simple consequence of the symmetry between 1 and 0 of the uniform distribution.

ii). It is a simple consequence of Theorem 2.2.1 and Example 1.7.9 in [17].

iii). Trivial.

For uniformly distributed inputs $\xi_1(\tau)$, when $t \to \infty$ we have

$$\epsilon(t,1) := \langle P(\xi_1(t+1) \leq \xi_1(tt)|\mathcal{F}_t)\rangle = \frac{1}{t} \tag{5}$$

Proof By definition of $\epsilon(t,1)$ (Eq. (5)) and Eq. (4) we get

$$\epsilon(t,1) = \langle \xi_1(tt)\rangle = \int_0^\infty xte^{-tx}dx = \frac{1}{t} \tag{6}$$

Now we turn our attention to more general case: the input signals are continuously distributed random variables. By this we mean that the Radon-Nikodyn derivative of the input distribution is absolutely continuous with respect to the Lebesgue measure. Denote the density

$$f(x) = dP/dx$$

From the definition of $\epsilon(t,1)$ (Eq. (5)) we see that $\epsilon(t,1) = \int_0^{\xi_1(tt)} f(x)dx$. Define a transformation $Y : I\!R^1 \to I\!R^1$ by $Y(x) = \int_0^x f(u)du$ then Eq. (4) becomes

$$\epsilon(t,1) = \int_{Y(0)}^{Y(\xi_1(tt))} dY(x) \tag{7}$$

Since the function is Y is a nondecreasing function we conclude that $Y(\xi_1(tt)) \geq Y(\xi_1(t(t-1))) \geq \cdots \geq Y(\xi_1(tk)) \geq ...,\ k < t-1$ which yields

If ξ_1 is a continuously distributed random variable we have

$$\epsilon(t,1) = 1/t$$

Lemma 3 also gives rise to a transparent and elementary proof of the universal property of the generalisation errors of the simple perceptron in one dimensional case: $\epsilon(t,1)$ is independent of the distribution of inputs; $\epsilon(t,1) = 1/t$ for whatever continuously distributed inputs. With the help of lemmas above and the assumption below we consider the case of two dimensional inputs.

Assumption 2: We suppose that $\xi_2(\tau) \sim 1/2(\delta_0 + \delta_1)$, *i.e. inputs signals are drawn from two lines* $y = 0$ *and* $y = 1$.

Suppose that $\xi_1(\tau)$ is continuously distributed. As $t \to \infty$ we have the following assertion

$$\epsilon(t,2) = 2\,\frac{1}{t} \tag{8}$$

Proof The following identity is a basic one which indicates that when $\xi_2(tt) \neq \xi_2(t(t-1))$ $\Omega(t)$ is simply the region on the left side of the line passing through $(\xi_1(tt), \xi_2(tt))$ and $(\xi_1(t(t-1)), \xi_2(t(t-1)))$; when $\xi_2(tt) = \xi_2(t(t-1))$ but $\xi_2(t(t-2)) \neq \xi_2(t(t-1))$ then $\Omega(t)$ is the region on the left side of the line passing through $(\xi_1(tt), \xi_2(tt))$ and $(\xi_1(t(t-2)), \xi_2(t(t-2)))$; $\cdots$

$$\begin{aligned}\epsilon(t,2) = \langle [&P(\xi(t+1) \in \Omega(t) | \xi_2(tt) \neq \xi_2(t(t-1))))) I_{\{\xi_2(tt)\neq\xi_2(t(t-1))\}} \\ &+P(\xi(t+1) \in \Omega(t) | \xi_2(tt) = \xi_2(t(t-1)) \neq \xi_2(t(t-2))) \\ &\quad \cdot I_{\{\xi_2(tt)=\xi_2(t(t-1))\neq\xi_2(t(t-2))\}} \\ &+P(\xi(t+1) \in \Omega(t) | \xi_2(tt) = \xi_2(t(t-1)) = \xi_2(t(t-2)) \neq \xi_2(t(t-3))) \\ &\quad \cdot I_{\{\xi_2(tt)=\xi_2(t(t-1))=\xi_2(t(t-2))\neq\xi_2(t(t-3))\}} \\ &+\cdots]\rangle\end{aligned} \tag{9}$$

where I is the indicator function. Therefore to obtain an exact expression of $\epsilon(t,2)$ it suffices for us to consider each term in Eq. (9). In fact we see that

$$\begin{aligned}&P(\xi(t+1) \in \Omega(t) | \xi_2(tt) = \xi_2(t(t-1)) = \cdots = \xi_2(t(t-k)) \neq \xi_2(t(t-k-1))) \\ &= \frac{1}{2}[\int_0^{\xi_2(tt)} dx + \int_0^{\xi_2(t(t-k-1))} dx] \\ &= \frac{1}{2}[\xi_1(tt) + \xi_1(t(t-k-1))]\end{aligned} \tag{10}$$

Note that $1/2^k = \langle\langle I_{\{\xi_2(tt)=\xi_2(t(t-1))=\cdots=\xi_2(t(t-k))\neq\xi_2(t(t-k-1))\}}\rangle\rangle$, together with Eq.(10) we derive that

$$\begin{aligned}&\langle [P(\xi(t+1) \in \Omega(t) | \xi_2(tt) = \xi_2(t(t-1)) = \cdots = \xi_2(t(t-k)) \neq \xi_2(t(t-k-1))) \\ &\quad \cdot I_{\{\xi_2(tt)=\xi_2(t(t-1))=\cdots=\xi_2(t(t-k))\neq\xi_2(t(t-k-1))\}}]\rangle \\ &= \frac{1}{2}[\frac{1}{2^k}(\langle\xi_1(tt)\rangle + \langle\xi_1(t(t-k-1))\rangle)]\end{aligned} \tag{11}$$

Substituting Eq. (11) into Eq. (9), in terms of Lemma 3 we finally conclude that

$$\epsilon(t,2) = \frac{1}{2^2}(\langle\xi_1(tt)\rangle + \langle\xi_1(t(t-1))\rangle) + \frac{1}{2^3}(\langle\xi_1(tt)\rangle + \langle\xi_1(t(t-2))\rangle) + \cdots \tag{12}$$

$$= \frac{1}{2}\langle\xi_1(tt)\rangle + \frac{1}{2^2}\langle\xi_1(t(t-1))\rangle + \frac{1}{2^3}\langle\xi_1(t(t-k-2))\rangle + \cdots \tag{13}$$

$$= \frac{1}{2}\frac{1}{t} + \frac{1}{2^2}\frac{2}{t} + \frac{1}{2^3}\frac{3}{t} + \cdots \tag{14}$$

$$= 2 \cdot \frac{1}{t} \tag{15}$$

Eq. (13) is the key identity of our approach which claims that $\epsilon(t,2)$ is a linear combination of extremes under assumption 1 and assumption 2. The identity enables us to obtain, in conjunction with extreme value theory, an exact expression for generalisation errors of the simple perceptron. It is readily seen that all conclusions in Theorem 1 is valid when $\xi_2(\tau) \sim p\delta_0 + q\delta_1, p > 0, q > 0, p+q = 1$.

To confirm our theoretical approach above: coefficient of the generalisation error of the simple perceptron is independent of inputs, here we include a numerical simulation to estimate the generalisation error. Let *both* $\xi_1(\tau), \xi_2(\tau)$ be i.i.d. and uniformly distributed over $[0,1]$. Fig. 3 shows the numerical results with 10000 times simulations for each $t = 100, 200, \cdots, 10000$. In [9] numerical simulations are presented for a variety of input distributions included in NAG library. Numerical results underpin our theoretical approach: the exact form of the generalisation error of the simple perceptron can be obtained under assumption 1 and assumption 2.

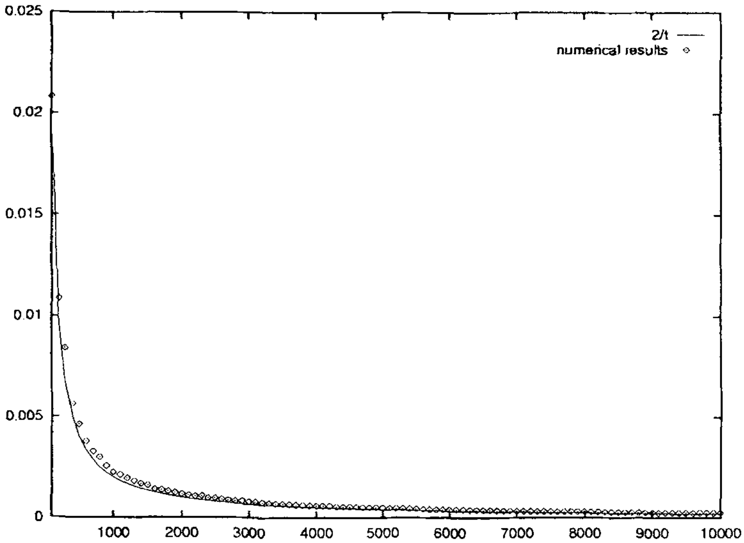


Fig. 2. Numerical simulations of $\epsilon(t, 2)$ when inputs $(\xi_1(\tau), \xi_2(\tau))$ are i.i.d uniformly distributed random variables. $\epsilon(t, 2)$ for $t = 100, 200, 300, \cdots, 10000$ are numerically calculated.

Remark 1 Surprisingly, our numerical and theoretical results are both different from the results obtained in terms of the replica trick approach in which it is estimated that $\epsilon(t, m) = 0.62m/t$. The deviation can be understood from the following two reasons: firstly the replica trick approach as we already pointed at the beginning of the paper is valid only when m tends to infinity in proportion to t; secondly the behaviour of extreme value also changes substantially when k is proportion to t, see for example [17]. However when m is small this effect will not play a role in our estimation since in Eq. (13) the term with large k is quite small already. But when $m \to \infty$ in proportion to t we have to take into account this effect in Eq. (13) .

Remark 2 It is easily seen that the approach above can be generalised to any dimensional case: assume $\xi_2(\tau)$ are distributed subjected to $1/m(\delta_{(x_2=0, x_3=0, \cdots, x_m=0)} +$

$\delta_{(x_2=1,x_3=0,\cdots,x_m=0)} + \cdots + \delta_{(x_2=0,x_3=0,\cdots,x_m=1)})$–the simplest distribution embodying geometrical structure of m dimension, the problem to find the generalisation error is thus reduced to calculate probabilities like in Eq. (11)–Eq. (17). We found that(see Remark 1)

$$\epsilon(t,m) = \begin{cases} \dfrac{1}{t} & \text{if } m = 1 \\ \dfrac{(m-1)!}{(m-1)^{(m-1)}}(\dfrac{m}{2}+1)\dfrac{1}{t} & \text{otherwise} \end{cases} \tag{16}$$

A detailed proof can be found in our full paper[9].

5 Discussion

There remain a lot questions for further investigation. For example, a challenging problem is to generalise our approach to consider algorithms like the BP algorithm etc. [1, 8, 10, 11, 12]. It is promising: to replace the line we considered in this paper by a curve reflecting the nonlinearity of the BP and the curve is determined by a few(more than two in the two dimensional case) extreme values of input signals; to take a similar approach as we developed here, we would expect to obtain a learning curve for the BP algorithm.

In summary our approach reported in this paper opens up new possibilities for rigorous analyses of generalization errors which reflect intricate nonlinear properties underlying most learning algorithms in neural networks.

Acknowledgement. I thank David Brown for valuable comments on the manuscript.

References

1. Albeverio, S., Feng, J., and Qian, M.(1995), The role of noises in neural networks, *Phys. Rev. E.*, **52**, 6593-6606.
2. Amari, S., Murata, N., and Ikeda, K. (1995), Statistical theory of learning curves, in: Oh, J., Kwon, Ch., and Chao, S.(eds), *Neural Networks: The Statistical Mechanics Perspective* , 3-17.
3. Baum, E.B.(1990), The perceptron algorithm is fast for nonmalicious distribution, *Neural computation*, 2, 248.
4. Baum, E.B., and Haussler, D.(1989), What size net gives valid generalization, *Neural computation*, 4, 151-160.
5. Cohn, D., and Tesauro, G.(1992), How tight are the Vapnik-Chervonenkis bounds, *Neural Computation*, 4, 249-269.
6. Engel, A., and den Broeck, C.V.(1993), Statistical mechanics calculation of Vapnik Chervonenkis bounds for perceptrons, *J. Phys*, **26** 6893-6914.
7. Feng, J.(1997), Behaviours of spike output jitter in the integrate-and-fire model. *Phys. Rev. Letters* (in press).
8. Feng, J.(1997), Lyapunov functions for neural nets with nondifferentiable input-output characteristics, *Neural Computation*, **9**, 45-51.

9. Feng, J. (1997), Generalisation error of the simple perceptron, (preprint).
10. Feng, J., and Hadeler, K. P.(1996), Qualitative behaviors of some simple neural networks, *J. Phys. A*, **29**, 5019-5033.
11. Feng, J., Pan, H., and Roychowdhury, V. P.(1996), On neurodynamics with limiter function and Linsker's developmental model, *Neural Computation*, **8**, 1003-1019.
12. Feng, J., and Tirozzi, B.(1995), The SLLN for the free-energy of the Hopfield and spin glass model, *Helvetica Physica Acta*, **68**, 365-379.
13. Galambos, J.(1984), *Introductory Probability Theory*, Marcek Dekker, INC., New York, 164-168.
14. Gray, M.S., Lawrence, D.T., Golomb, B.A., and Sejnowski, T.J.(1995), A perceptron reveals the face of sex, *Neural Computation* **7**, 1160-1164.
15. Haussler, D., Kearns, M., and Shapire, R.(1991), Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension, *Proc. 4th Ann. Workshop on computational Learning Theory*, Morgan Kaufmann, San Mateo, CA, 61-74.
16. Haussler, D., Littlestone, N., and Warmuth, K.(1988), Predicting $\{0,1\}$ functions on randomly drawn points, *Proc. COLT'88*, Morgan Kaufmann, San Mateo, CA, 280-295.
17. Leadbetter, M.R., Lindgren, G., and Rootzén, H.(1983), *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York, Heidelberg, Berlin.
18. Levin, E., Tishby, N., and Solla, S.A.(1990), A statistical approach to learning and generalization in layered neural networks, *Proceeding of the IEEE*, **78**(10), 1568-1574.
19. Murata, N., Yoshizawa, S., and Amari, S.(1994), Network information criterion-determinate the number of hidden units for an artificial neural network model, *IEEE Trans. NN*, **6**, 865-872.
20. Newman, C., and Stein, D.L.(1996), Non-mean-field behavior of realistic spin glass, *Physical Review Letter* **76**(3), 515-518.
21. Opper, M., and Haussler, D.(1991), Calculation of the learning curve of Bayes optimal classification algorithm for learning perceptron with noise, *Proceedings of the Fourth Annual Workshop on Computer Learning Theory*, 75-87.
22. Opper, M., and Haussler, D.(1995), Bounds for predictive errors in the statistical mechanics of supervised learning, *Physical Review Letter* **75**, 3772-2775.
23. Seung, H.S., Sompolinsky, H., and Tishbby, N.(1992), Statistical mechanics of learning from examples, *Physical Review A*, **45**, 6056-6091.
24. Vapnik, V.N., and Chervonenkis, A.Y.(1971), On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probab. and its Appl.* **16**(2), 264-280.
25. Vapnik, E., Levin, E., and LeCun, Y.(1994), Measuring the VC dimension of a learning machine, *Neural Computation*, **5**, 851-876.
26. Watkin, T.L.H., Rau, A., and Biehl, M.(1993), The statistical mechanics of learning a rule, *Rev. Mod. Phys.*, **65**, 499-556.
27. Yamanishi, K.(1991), A loss bound model for on-line stochastic prediction strategies, *Proceeding of the Fourth Annual Workshop on Computer Learning Theory*, 290-302.