Learning with noise in a linear perceptron

1992 J. Phys. A: Math. Gen. 25 1119

(http://iopscience.iop.org/0305-4470/25/5/019)

J. Phys. A: Math. Gen. 25 (1992) 1119–1133. Printed in the UK

Learning with noise in a linear perceptron

Anders Krogh†
The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen, Denmark

Received 2 April 1991, in final form 4 November 1991

Abstract. The learning of a set of p random patterns in a linear perceptron is studied in the limit of a large number (N) of input units with noise on the weights, inputs and output. The problem is formulated in continuous time as a Langevin equation, and the first task is to evaluate the response or Green's function for the system. White noise on the output is shown to correspond to spatially correlated weight noise acting only in a subspace of the weight space. It is shown that the input noise acts as a simple weight decay with a size proportional to the load parameter $\alpha = p/N$. With no weight decay, the relaxation time diverges at $\alpha = 1$. With a weight decay it becomes shorter, and finite for $\alpha = 1$, but at the cost of a larger asymptotic learning error that is is found analytically. It is shown that a small weight decay decreases the effect of noise on the weights or outputs.

1. Introduction

In recent years there has been much interest in neural network models because of their ability to learn from examples. For some of these models it has been possible to calculate, analytically, certain quantities like the capacity [1–3]. However little analytical work has been done on the dynamics of the learning process i.e. asking questions about learning time and the trade-offs between learning time and accuracy.

In this paper delta rule or adaline learning is considered in the simple linear perceptron without hidden units, and earlier [4–7] results on asymptotic learning times and capacity in the presence of noise are extended. In a previous paper noise on the connections was studied, and here I elaborate on those results and carry out the same kind of analysis for the case of noisy inputs and outputs. In our previous paper we concentrated mostly on 'constrained learning', i.e. learning where the size of the weight vector is constrained to a particular length. In this paper the emphasis is on unconstrained learning.

The network is a standard linear perceptron with N input units. Output units can always be treated separately, so it is sufficient to study *one* linear unit receiving inputs $\xi \in R^N$ and producing an output

$$V = N^{-1/2} \sum_{i=1}^{N} w_i \xi_i \tag{1}$$

† Present address: Computer and Information Sciences, University of California, Santa Cruz, CA 95064, USA.

0305-4470/92/051119+15$04.50

with w_i the synaptic weights. It is assumed that a training set of p examples $(\xi^\mu, \zeta^\mu), \mu = 1, \ldots, p$, is given, and the aim of the learning procedure is to find weights such that

$$V^\mu = N^{-1/2} \sum_{i=1}^{N} w_i \xi_i^\mu = \zeta^\mu. \tag{2}$$

The learning process is formulated as a gradient descent minimization of an energy or cost function, which is usually taken to be the squared error of the output, $(\zeta - V)^2$. A term that penalizes large weights, $\frac{1}{2}\lambda \sum_i w_i^2$, is added to make it possible to limit the size of the weights, so the cost function reads

$$E = \frac{1}{2} \sum_\mu (\zeta^\mu - V^\mu)^2 + \frac{1}{2} \lambda \sum_i w_i^2. \tag{3}$$

The change of weights is proportional to the negative gradient of E, and in continuous time this becomes

$$\dot{w}_i = -\gamma_0 \frac{\partial E}{\partial w_i} = \gamma_0 \left(B_i - \sum_j A_{ij} w_j - \lambda w_i \right) \tag{4}$$

where

$$B_i = N^{-1/2} \sum_\mu \zeta^\mu \xi_i^\mu \tag{5}$$

$$A_{ij} = \frac{1}{N} \sum_\mu \xi_i^\mu \xi_j^\mu. \tag{6}$$

It is assumed that the input patterns as well as the targets are random. Note that 'batch update' is assumed, i.e. all patterns are presented before the weights are changed.

Three kinds of dynamic noise will be considered.

(i) Noise on the inputs. White noise $\epsilon_i^\mu(t)$ added to the input patterns ξ_i^μ. This corresponds to, e.g. noisy input lines in an electronic implementation.

(ii) Noise on the weights. White noise $\eta_i(t)$ added to the learning equation (4),

$$\dot{w}_i \propto -\frac{\partial E}{\partial w_i} + \eta_i(t)$$

corresponding to noise in the weight update procedure. This is the most commonly studied form of noise (see, e.g., [10]) and it leads in equilibrium to a Gibbs distribution.

(iii) Noise on the outputs. White noise $f^\mu(t)$ is added to the target similar to the input noise. It enters the learning equation only through the error $\zeta^\mu - V^\mu + f^\mu(t)$, and can therefore be viewed as noise in the output determination, noise on the targets (noisy teacher), or noise in the error signal. This kind was studied by Der [7] and the results will be reviewed and expanded.

All three kinds of noise are relevant for implementations of neural networks.

2. The response function

In Fourier transformed form the solution to (4) can be written compactly as

$$w_i(\omega) = G_0^{-1}(\omega) \sum_j g_{ij}(\omega) w_j^0(\omega). \tag{7}$$

Here the response (or Green's) function was introduced,

$$\mathbf{g}(\omega) = [\mathbf{A} + (\lambda - i\omega/\gamma_0)\mathbf{I}]^{-1} \tag{8}$$

its non-interacting limit,

$$G_0(\omega) = \frac{1}{\lambda - i\omega/\gamma_0} \tag{9}$$

and

$$w_i^0(\omega) = G_0(\omega)(B_i 2\pi\delta(\omega) + c_i). \tag{10}$$

The c_i are constants depending on the initial condition. It is assumed that $w_i(t) = 0$ for $t < 0$.

Note that in the static limit, $t \to \infty$, $\dot{w}_i = 0$, so the solution to (4) is just

$$w_i(t \to \infty) = \sum_j g_{ij}(\omega = 0) B_j. \tag{11}$$

Here and in the rest of the paper $w_i(t = 0) = 0$ is assumed. The limit $\lambda \to 0$ corresponds to the pseudo-inverse solution to the learning problem [8, 9]. If λ is very large $\mathbf{g}$ is completely dominated by λ, and $w_i \simeq \lambda^{-1} B_i$ which corresponds to Hebbian synapses (see (5)). Therefore λ can be used as a parameter to interpolate between the pseudo-inverse and the Hebbian solutions.

Most quantities of interest can be calculated from the average response function, $G_{ij}(\omega) \equiv [g_{ij}(\omega)]_\xi$ which does not depend on the targets. $[\cdot]_\xi$ means the average over the random input patterns ξ^μ. To find the average one begins by expanding (8),

$$G_{ij}(\omega) = \left[G_0(\omega) - G_0^2(\omega) A_{ij} + G_0^3(\omega) \sum_k A_{ik} A_{kj} - \cdots \right]. \tag{12}$$

This equation can also be found by iteration of the Dyson equation, which is derived directly from (8) by multiplying both sides by $G_0[\mathbf{A} + (\lambda - i\omega/\gamma_0)\mathbf{I}]$

$$\mathbf{g} = G_0 - G_0 \mathbf{A} \mathbf{g}. \tag{13}$$

The average can be found by diagrammatic methods as described in [4]. The self-energy, Σ, is defined by

$$G_{ij}^{-1}(\omega) = G_0^{-1}(\omega)\delta_{ij} + \Sigma_{ij}(\omega). \tag{14}$$

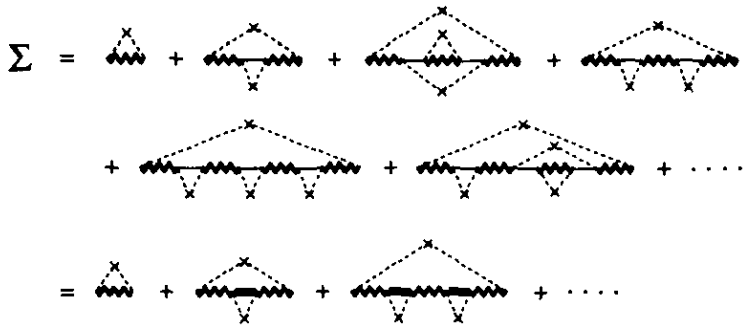

Figure 1. Diagrams for the self-energy. The wiggly line represents **A**, the single line $-G_0$, and the double line $-G$. The pairing (and averaging) of the ξ is shown by the dotted lines and the x. The ends connected by these lines have to have the same pattern and unit index. The last line shows how the self-energy can be drawn with 'dressed' diagrams.

It is assumed that the components of the ξ are random and independent with zero mean and variance σ^2,

$$[\xi_i^\mu]_\xi = 0 \qquad \text{and} \qquad [\xi_i^\mu \xi_j^\nu]_\xi = \delta_{ij}\delta_{\mu\nu}\sigma^2. \tag{15}$$

The diagrams for the self-energy that survive in the $N \to \infty$ limit are shown in figure 1. (In [4] it is shown how one handles the general case of non-zero mean.)

There are only diagonal terms in the expansion for Σ, proportional to

$$\left[N^{-q}\sum_\mu \sum_{i_1 i_2 \ldots i_{q-1}} \xi_i^\mu \xi_{i_1}^\mu \ldots \xi_{i_{q-1}}^\mu\right]_\xi = \alpha\sigma^{2q} \tag{16}$$

and this diagonal element (denoted by Σ) becomes

$$\Sigma = \sigma^2\alpha(1 - \sigma^2 G + (\sigma^2 G)^2 - (\sigma^2 G)^3 + \cdots) = \frac{\sigma^2\alpha}{1+\sigma^2 G}. \tag{17}$$

G_{ij} is also clearly diagonal, and the diagonal element is given by

$$G^{-1}(\omega) = G_0^{-1}(\omega) + \Sigma(\omega) = \lambda - \frac{\mathrm{i}\omega}{\gamma_0} + \frac{\sigma^2\alpha}{1+\sigma^2 G}. \tag{18}$$

Putting $z = \lambda - \mathrm{i}\omega/\gamma_0$ and solving the equation yields G as a function of z,

$$G(z) = \frac{1 - \alpha - z/\sigma^2 \pm \sqrt{(z-z_+)(z-z_-)/\sigma^4}}{2z}. \tag{19}$$

One has to choose the '+' branch in order to get a causal solution: a $1/z$ behaviour in the large z limit. Here

$$z_\pm = -(1 \pm \sqrt{\alpha})^2\sigma^2 \tag{20}$$

was introduced.

It is often convenient to know the average eigenvalue spectrum of the matrix **A**, for instance in order to integrate the learning error as a function of time. If A_r are the eigenvalues of **A** the average spectrum can be written as

$$\rho(x) = \frac{1}{N}\sum_r \delta(x - A_r) = \frac{1}{\pi N}\sum_r \lim_{\delta\to 0} \mathrm{Im}\left[(x + \mathrm{i}\delta - A_r)^{-1}\right]_\xi$$
$$= \frac{1}{\pi}\lim_{\delta\to 0}\mathrm{Im}\frac{1}{N}\,\mathrm{tr}\,\mathbf{G}(z = -x - \mathrm{i}\delta). \tag{21}$$

From (19) one finds†

$$\rho(x) = (1-\alpha)\Theta(1-\alpha)\delta(x) + \frac{\sqrt{-(x+z_+)(x+z_-)/\sigma^4}}{2\pi x}. \tag{22}$$

It is implicit that the last term only contributes in the real regime between the roots $-z_-$ and $-z_+$. The spectrum consists of a peak—the delta function—at $x = 0$, and a semicircle deformed by $1/x$ between $-z_-$ and $-z_+$.

It is clear that σ^2 just scales the size of the eigenvalues, and can only play this role of a scaling factor. Therefore, without loss of generality, it will be assumed that $\sigma^2 = 1$ in much of the rest of the paper.

3. Learning with noise

In this section the three kinds of noise mentioned in the introduction will be introduced.

3.1. Weight noise

The effect of noise on the weights has been investigated [4]. This noise is represented by adding a white noise $\eta_i(t)$ to the learning equation (4)

$$\dot{w}_i = \gamma_0\left(B_i - \sum_j A_{ij}w_j - \lambda w_i\right) + \eta_i(t). \tag{23}$$

The noise has correlation

$$\langle \eta_i(t)\eta_j(t')\rangle_T = 2T\gamma_0\delta_{ij}\delta(t-t') \tag{24}$$

where T is a 'temperature' or noise level for the weight noise, and $\langle\cdot\rangle_T$ denotes an average over this noise. In Fourier transformed form the solution (7) to the learning equation now reads

$$w_i(\omega) = \sum_j g_{ij}(\omega)\left(G_0^{-1}(\omega)w_j^0(\omega) + \frac{\eta_j(\omega)}{\gamma_0}\right). \tag{25}$$

When w_i is averaged over noise the noise term will simply disappear; it enters only in terms with $\langle w_i^2\rangle_T$. Therefore relaxation times etc, are independent of T, as will be seen in the next section.

The equal-time correlation function is defined as

$$c_{ij} \equiv \langle w_i(t)w_j(t)\rangle_T - \langle w_i(t)\rangle_T\langle w_j(t)\rangle_T. \tag{26}$$

There exists an important relation (derived in appendix A) between this and the response function

$$\mathbf{c} = T\mathbf{g}(\omega = 0). \tag{27}$$

† This spectrum can also be calculated by the replica method, see [5].

3.2. Output noise

Adding white noise $f^\mu(t)$ to the target, $\zeta^\mu \longrightarrow \zeta^\mu + f^\mu(t)$, gives a learning equation of the same form with effective noise

$$\hat{\eta}_i(t) = N^{-1/2} \sum_\mu \xi_i^\mu f^\mu(t). \tag{28}$$

Since this noise is additive (and $\langle \hat{\eta}_i(t) \rangle = 0$) the noise average of w_i is independent of the noise, as with weight noise, and the relaxation times will be the same.

In this case the noise is projected onto the subspace spanned by the patterns, so it does not act in the orthogonal subspace. The variance of the noise is

$$\langle \hat{\eta}_i(t) \hat{\eta}_j(t') \rangle_T = A_{ij} \langle f^\mu(t) f^\mu(t') \rangle_T = 2T\gamma_0 A_{ij} \delta(t - t'). \tag{29}$$

In each eigendirection of $\mathbf{A}$ the noise is therefore white, but multiplied by the corresponding eigenvalue of $\mathbf{A}$. This spatially correlated noise changes the fluctuations in the system if compared to the uncorrelated weight noise. The difference is clearly seen in the different form of the fluctuation-response relation, which now reads (see appendix A)

$$\mathbf{c} = T\mathbf{A}\mathbf{g}(\omega = 0) = T(\mathbf{I} - \lambda \mathbf{g}(\omega = 0)). \tag{30}$$

3.3. Input noise

The analysis is now extended to the case of noise on the input units, i.e. in addition to the actual value of the input pattern there is noise. If white noise like (24) is assumed, it turns out that the delta function would make the self-energy blow up. Therefore the noise must have a finite size at $t = 0$, and the following form of the correlation function is chosen

$$\langle \epsilon_i^\mu(t) \epsilon_j^\nu(t') \rangle_\gamma = \gamma \delta_{ij} \delta_{\mu\nu} e^{-\Delta|t-t'|}. \tag{31}$$

It will be assumed that Δ is large. The level of this noise is called γ, and averages are denoted $\langle \cdot \rangle_\gamma$.

The noise is included in the B_i (5) and A_{ij} (6) by defining

$$B_i^\epsilon(t) = N^{-1/2} \sum_\mu \zeta^\mu (\xi_i^\mu + \epsilon_i^\mu(t)) \tag{32}$$

$$A_{ij}^\epsilon(t) = \frac{1}{N} \sum_\mu (\xi_i^\mu + \epsilon_i^\mu(t))(\xi_j^\mu + \epsilon_j^\mu(t)). \tag{33}$$

The dynamics will then be governed by the same equation (4) with these new time-dependent A and B. Fourier transforming (4) now yields

$$w_i(\omega) = w_i^0(\omega) - G_0(\omega) A_{ij}^\epsilon(\omega - \omega') w_j(\omega'). \tag{34}$$

In this and later equations there are implicit sums on indices appearing twice in a term, and integrals (divided by 2π) over doubly-occurring ω. The form of this equation is similar to the one obtained without input noise. Iteration gives

$$\begin{aligned} w_i(\omega) = [&G_0(\omega)\delta_{ij}\delta(\omega - \omega') - G_0(\omega) A_{ij}^\epsilon(\omega - \omega') G_0(\omega') + G_0(\omega) A_{ik}^\epsilon(\omega - \omega'') \\ &\times G_0(\omega'') A_{kj}^\epsilon(\omega'' - \omega') G_0(\omega') - \cdots] G_0^{-1}(\omega') w_j^0(\omega'). \end{aligned} \tag{35}$$

Inside the square brackets is the unaveraged response function, so in this case the average one is

$$G_{ij}(\omega-\omega') = \left\langle \left[G_0(\omega)\delta_{ij} - G_0(\omega)A^{\epsilon}_{ij}(\omega-\omega')G_0(\omega') \right.\right. \\ \left.\left. + G_0(\omega)A^{\epsilon}_{ik}(\omega-\omega'')G_0(\omega'')A^{\epsilon}_{kj}(\omega''-\omega')G_0(\omega') - \cdots \right]_{\xi} \right\rangle_{\gamma}. \quad (36)$$

(No integral on ω' here.) Here only the case of uncorrelated patterns where the Green's function is diagonal is considered, although extension to biased patterns should be easy (see [4]).

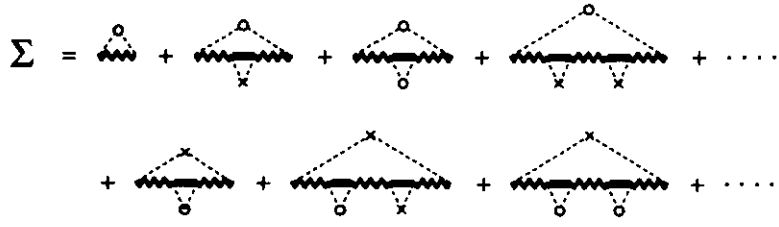

Figure 2. Some of the diagrams for the new self-energy. Two ends connected by a dotted line with a circle means pairing of two ϵ and ones connected with dotted lines with a cross means pairing of two ξ as before. Only the first of these diagrams contributes to the lowest order in $1/\Delta$. The self-energy diagrams in figure 1 still contributes in this case, but they are not shown here.

To give a contribution the ϵ have to be paired, just like the patterns, giving rise to self-energy diagrams of similar topology as the diagrams for the noiseless case, see figure 2. Luckily, it turns out that only the simplest of the new diagrams contributes to the self-energy to lowest order in $1/\Delta$. The Fourier transform of $e^{-\Delta|t|}$ is $2\Delta/(\Delta^2+\omega^2)$, so one finds the contribution from the first of the new diagrams to be

$$\int d\omega_1 \frac{1}{2\pi N}\sum_{\mu}\langle \epsilon_i^{\mu}(\omega_1)\epsilon_i^{\mu}(\omega-\omega'-\omega_1)\rangle_{\gamma} \\ = \frac{p}{2\pi N}\delta(\omega-\omega')\int d\omega_1 \frac{4\pi\gamma\Delta}{\Delta^2+\omega_1^2} = 2\pi\alpha\gamma\delta(\omega-\omega') \quad (37)$$

In appendix B other terms in the self-energy are calculated, and it is shown that their contribution is of relative order $1/\Delta$ or less. For large Δ the only additional term in the self-energy is then just $\alpha\gamma$, so this noise acts exactly like an additional weight decay term.

Including the $1/\Delta$ corrections modifies the magnitude of the weight decay a little and also effectively shrinks the variance σ^2 of the input patterns (equation (15)). See appendix B for details.

Input noise acts completely different from the two other kinds of noise described previously, and in the following it will simply be treated as a weight decay, $\lambda \neq 0$, without explicitly writing the term $\alpha\gamma$.

4. Relaxation times

$G(z)$ has a pole at $z = 0$ if $\alpha < 0$ and a branch cut along the real axis between z_+ and z_- given by (20). The pole has the following meaning. For $\alpha < 1$ the pattern vectors $\boldsymbol{\xi}^\mu$ span a p-dimensional subspace of the input space which will be called the pattern subspace. If $\lambda = 0$ the dynamics (4) acts only in this subspace. Therefore there will be a non-relaxing component of w in the complementary subspace. This gives rise to the pole at $z = \omega = 0$. For $\lambda > 0$ the part of w outside the pattern subspace will decay exponentially with exponent λ. In either case the pole does not describe the interesting part of the dynamics (i.e. that in the pattern subspace), so it is subtracted from G.

For $\alpha < 1$ and small z, $G(z) \simeq (1 - \alpha)/z$, and a new G is defined

$$\hat{G}(z) = \begin{cases} G(z) - (1-\alpha)/z & \text{for } \alpha \leqslant 1 \\ G(z) & \text{for } \alpha > 1. \end{cases}$$

Then $\hat{G}$ can be written as

$$\hat{G}(z) = \frac{-|1-\alpha| - z + \sqrt{(z - z_+)(z - z_-)}}{2z} \tag{38}$$

for *both* $\alpha > 1$ *and* $\alpha < 1$.

The characteristic relaxation time can be found from the response function (see [4])

$$\tau = -\frac{1}{\gamma_0 \hat{G}(z)} \left.\frac{\partial \hat{G}(z)}{\partial z}\right|_{z=\lambda}. \tag{40}$$

Differentiating (39) gives, after a little work,

$$\gamma_0 \tau = \frac{1}{2\lambda}\left[1 + \frac{\lambda - \sqrt{z_+ z_-}}{\sqrt{(\lambda - z_+)(\lambda - z_-)}}\right]. \tag{41}$$

In the limit of no decay ($\lambda = 0$) the previous result [4, 5] is recovered,

$$\gamma_0 \tau = \begin{cases} 1/(1-\alpha)^2 & \text{for } \alpha < 1 \\ \alpha/(1-\alpha)^2 & \text{for } \alpha > 1. \end{cases} \tag{42}$$

There is a critical slowing down as the critical point ($\alpha = 1$) is approached, and right at the critical point $G(z) \propto z^{-1/2}$, leading to a $t^{-1/2}$ error decay instead of an exponential one.

Figure 3 shows the relaxation time as a function of α for $\lambda = 0.2$. The qualitative behaviour is the same for other values of λ, but the relaxation time decreases with λ (at the cost of a higher learning error, which will be calculated in the next section).

Instead of considering this average relaxation time, one could instead look at the longest relaxation time in the system. This is given by the inverse of the smallest eigenvalue of $\mathbf{A} + \lambda \mathbf{I}$. The smallest eigenvalue of $\mathbf{A}$ is $-z_-$, giving a relaxation time

$$\gamma_0 \tau' = \frac{1}{\lambda + (1 - \sqrt{\alpha})^2}. \tag{43}$$

This is also shown in figure 3. If the weight decay is chosen equal to $\gamma\alpha$, corresponding to input noise, the relaxation time follows a curve like the dotted one in figure 3. The phase transition smears out because of the noise, but it is interesting to note that what is left of it (the peak in τ') moves to smaller α. By inserting $\lambda = \gamma\alpha$ into (43) and using $\partial\tau'/\partial\alpha = 0$ one finds the value of α that gives the longest relaxation time is $\alpha_{\max \tau'} = (1 + \gamma)^{-2}$.

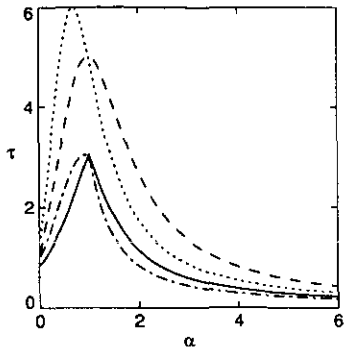

Figure 3. Relaxation times. The full curve is the average relaxation time τ, and the broken curve is the longest relaxation time τ', both for $\lambda = 0.2$. They peak at $\alpha = 1$. The chain curve is the average relaxation time for a weight decay of $\lambda = \gamma\alpha$ ($\gamma = 0.2$) corresponding to input noise of level γ. The dotted curve is the longest relaxation time for the same parameters.

5. Learning error

In this section I will calculate the size of the learning error. It is

$$\begin{aligned}\hat{E} &= \left[\langle \frac{1}{p}\sum_{\mu}(\zeta^{\mu} - V^{\mu})^2\rangle_T\right]_{\xi\zeta} \\ &= 1 + \frac{1}{p}\sum_{ij}\left[A_{ij}\langle w_i w_j\rangle_T\right]_{\xi\zeta} - \frac{2}{p}\sum_{i}\left[B_i\langle w_i\rangle_T\right]_{\xi\zeta}.\end{aligned} \tag{44}$$

Normalized this way, it is one if the output is zero (*tabula rasa*). It is assumed that $[\zeta^{\mu}\zeta^{\nu}]_{\zeta} = \delta_{\mu\nu}$.

The second term in the error can be found from the Langevin equation (23) by multiplying it by w_i and averaging,

$$\begin{aligned}0 &= \frac{1}{2}\sum_{i}\frac{\partial\langle w_i^2\rangle_T}{\partial t} = \sum_{i}\langle w_i\dot{w}_i\rangle_T \\ &= \gamma_0\sum_{i}\left(B_i\langle w_i\rangle_T - \sum_{j}A_{ij}\langle w_i w_j\rangle_T - \lambda\langle w_i^2\rangle_T\right) + \sum_{i}\langle\eta_i w_i\rangle_T.\end{aligned} \tag{45}$$

Using this to substitute for $A_{ij}\langle w_i w_j\rangle_T$ the error reads

$$\hat{E} = 1 - \frac{1}{p}\sum_{i}\left[B_i\langle w_i\rangle_T\right]_{\xi\zeta} - \frac{\lambda}{p}\sum_{i}\left[\langle w_i^2\rangle_T\right]_{\xi\zeta} + \frac{1}{\gamma_0 p}\sum_{i}\left[\langle\eta_i w_i\rangle_T\right]_{\xi\zeta}. \tag{46}$$

Since $\langle w_i\rangle_T = \sum_j g_{ij}(0)B_j$ in this limit (see (11)) the second of these terms becomes

$$\begin{aligned}\frac{1}{p}\sum_{i}\left[B_i\langle w_i\rangle_T\right]_{\xi\zeta} &= \frac{1}{pN}\sum_{i,j}g_{ij}\sum_{\mu,\nu}\left[\zeta^{\mu}\zeta^{\nu}\xi_i^{\mu}\xi_j^{\nu}\right]_{\xi\zeta} \\ &= \frac{1}{\alpha N}\operatorname{tr}\left[\mathbf{gA}\right]_{\xi} = \frac{1}{\alpha}(1 - \lambda G)\end{aligned} \tag{47}$$

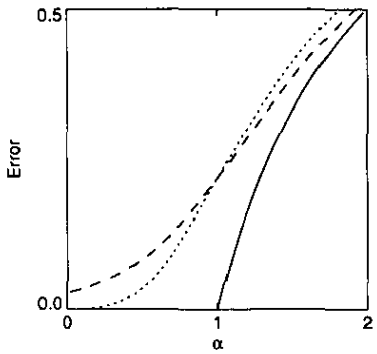

Figure 4. Asymptotic error as a function of α for $T = 0$. The full curve corresponds to $\lambda = 0$, the broken curve to $\lambda = 0.2$, and the dotted curve corresponds to $\lambda = \gamma\alpha$ ($\gamma = 0.2$). ($\gamma_0 = 1$ is assumed.)

where the Dyson equation (13) has been used in the last transformation. Here the response function written with no argument means it should be taken at $\omega = 0$, i.e. $G \equiv G(\omega = 0)$.

In [4] it is proven that if the targets ζ^μ are random

$$q \equiv \frac{1}{N}\sum_i \left[\langle w_i\rangle_T^2\right]_{\xi\zeta} = \frac{\partial}{\partial\lambda}(\lambda G). \tag{48}$$

q is part of the correlation function (26), so using the fluctuation-response relation (27) the third term in the error can be found,

$$\frac{\lambda}{p}\left[\langle w_i^2\rangle_T\right]_{\xi\zeta} = \frac{\lambda}{\alpha}\left(TG + \frac{\partial}{\partial\lambda}(\lambda G)\right). \tag{49}$$

To calculate the last term in (46) one multiplies the averaged Langevin equation (23) by $\langle w_i\rangle_T$,

$$0 = \sum_i \langle w_i\rangle_T\langle \dot{w}_i\rangle_T = \sum_i \gamma_0\Big(B_i\langle w_i\rangle_T - \sum_j A_{ij}\langle w_i\rangle_T\langle w_j\rangle_T - \lambda\langle w_i\rangle_T^2\Big). \tag{50}$$

Subtracting this from (45) and rearranging the terms gives

$$\sum_i \langle \eta_i w_i\rangle_T = \gamma_0\sum_{ij}(\lambda\delta_{ij} + A_{ij})\big(\langle w_i w_j\rangle_T - \langle w_i\rangle_T\langle w_j\rangle_T\big) = \gamma_0\sum_{ij} g_{ij}^{-1}c_{ij}. \tag{51}$$

Using again the fluctuation-response relation (27) $\mathbf{g}^{-1}\mathbf{c} = T\mathbf{I}$ and then

$$\frac{1}{\gamma_0 p}\sum_i \langle \eta_i w_i\rangle_T = T/\alpha. \tag{52}$$

Collecting all the terms the error (44) becomes

$$\hat{E} = \frac{\alpha - 1}{\alpha} - \frac{\lambda^2}{\alpha}\frac{\partial G}{\partial\lambda} + (1 - \lambda G)\frac{T}{\alpha}. \tag{53}$$

The asymptotic error as a function of α is depicted in figure 4 for some values of λ. By a comparison with figure 3 one observes that the asymptotic error increases when the learning time decreases.

Almost the same calculation applies for output noise. The part independent of the noise is unchanged. The only difference is that the fluctuation-response relation has to be replaced by (30). The result is

$$\hat{E} = \frac{\alpha - 1}{\alpha} - \frac{\lambda^2}{\alpha}\frac{\partial G}{\partial \lambda} + (1 - \frac{\lambda}{\alpha}(1 - \lambda G))T. \tag{54}$$

The simple linear T behaviour in these equations is due to the fact that the T-noise enters linearly in the Langevin equation. The $1/\alpha$ behaviour comes from the normalization of the error. If the error were normalized the same way as the original cost function E, the $1/\alpha$ would be absent.

If λ is sufficiently small $G(\lambda)$ can be expanded,

$$G(\lambda) \simeq \frac{1 - \alpha + |1 - \alpha|}{2\lambda} + \frac{1 + \alpha - |1 - \alpha|}{2|1 - \alpha|} - \frac{\alpha\lambda}{|1 - \alpha|^3}. \tag{55}$$

Then for $\lambda = 0$ the error is

$$\hat{E} = \begin{cases} T & \text{for } \alpha < 1 \\ \dfrac{\alpha - 1}{\alpha} + \dfrac{T}{\alpha} & \text{for } \alpha \geqslant 1 \end{cases} \tag{56}$$

with weight noise.

Keeping more terms in the expansion allows one to follow what happens for λ close to 0. In particular the derivative of $\hat{E}$ is always negative at $\lambda = 0$, meaning that the error *always decreases* close to $\lambda = 0$. Thus a weight decay is beneficial to learning. By putting the derivative equal to zero the minimum $\hat{E}$ can be found, and it turns out that the effect is relatively small, of order T^2 in this approximation at small T. But even if the error decrease is marginal it is worth remembering that a weight decay also decreases the learning time. Since a weight decay corresponds to input noise, one can say that weight noise and input noise act oppositely when balanced right.

The output noise model shows the same behaviour.

6. Conclusion

A *linear* 'network' consisting of *one unit* was studied, almost the simplest system one can imagine. Nevertheless a rich and interesting behaviour was found. There is a phase transition from perfect learning to imperfect learning at $\alpha = 1$, if $\lambda = 0$. For $\lambda \neq 0$ this transition is rounded, similarly to what happens in physical systems, giving a finite relaxation time even at $\alpha = 1$.

It was shown that the effect of noise in the input corresponds exactly to a weight decay of size $\gamma\alpha$. If the input noise has a finite time correlation ($\Delta < \infty$) the decay decreases slightly and the response function changes as if the input patterns had a more narrow distribution (to lowest order in $1/\Delta$). It is interesting to note that in the thermodynamic limit (infinite N) input noise can be completely compensated by a negative weight decay of same absolute size as the decay from the noise.

Noise on the output implies a pattern-dependent spatially correlated noise on the weights, acting only in the subspace of weight space spanned by the patterns.

Therefore the relaxation times are the same, and most other quantities like the asymptotic error behaves qualitatively the same for weight and output noise.

An interesting effect of weight decay was discovered: a small weight decay (or, equivalently, noise on the inputs) decreases the effect of weight or output noise. Since the weight decay at the same time decreases the learning time, it will often pay to learn with a small decay if the system is noisy.

Here only learning was studied. A very important issue is generalization, i.e. how well the network can imitate a teacher. If this teacher is another linear perceptron with fixed weights, this generalization ability can be calculated. This has been treated in [12] and is expanded in [13].

The hope is that the results from this simple system can guide us in the investigation of the more complicated and, from the point of view of applications, more interesting nonlinear networks.

Acknowledgment

I would like to thank John Hertz for his invaluable help and support.

Appendix A. Fluctuation-response relations

In the long time limit $w_i(\omega) = \gamma_0^{-1} \sum_j g_{ij}(\omega)\eta_j(\omega)$ (cf (25)), so the correlation function is

$$\begin{aligned} c_{ij}(\tau) &= \langle [w_i(t) - \langle w_i(t)\rangle_T][w_j(t+\tau) - \langle w_j(t+\tau)\rangle_T]\rangle_T \\ &= \int \frac{d\omega\, d\omega'}{(2\pi)^2} e^{-i(\omega+\omega')t - i\omega'\tau} \langle [w_i(\omega) - \langle w_i(\omega)\rangle_T][w_j(\omega') - \langle w_j(\omega')\rangle_T]\rangle_T \\ &= \gamma_0^{-2} \int \frac{d\omega\, d\omega'}{(2\pi)^2} e^{-i(\omega+\omega')t - i\omega'\tau} \sum_{lk} g_{ik}(\omega) g_{jl}(\omega') \langle \eta_k(\omega)\eta_l(\omega')\rangle. \end{aligned} \tag{A1}$$

For both weight noise and output noise the last noise average will produce a $2\pi\delta(\omega + \omega')$, so

$$c_{ij}(\tau) = \gamma_0^{-2} \int \frac{d\omega}{2\pi} e^{i\omega\tau} \sum_{lk} g_{ik}(\omega) g_{jl}(-\omega) \langle \eta_k(\omega)\eta_l(-\omega)\rangle \tag{A2}$$

or

$$c_{ij}(\omega) = \gamma_0^{-2} \sum_{lk} g_{ik}(\omega) g_{jl}(-\omega) \langle \eta_k(\omega)\eta_l(-\omega)\rangle. \tag{A3}$$

With the uncorrelated noise on the weights given by (24) $\langle \eta_k(\omega)\eta_l(-\omega)\rangle = 2\gamma_0 T \delta_{kl}$, so

$$c_{ij}(\omega) = 2\gamma_0^{-1} T \sum_k g_{ik}(\omega) g_{jk}(-\omega). \tag{A4}$$

The product $\mathbf{g}(\omega)\mathbf{g}(-\omega)$ can be found from the series expansion of $\mathbf{g}$ (12) (implicit sums on repeated indices):

$$\begin{aligned} g_{ik}(\omega)g_{jk}(-\omega) &= [G_0(\omega)\delta_{ik} - G_0^2(\omega)A_{ik} + G_0^3(\omega)A_{il}A_{lk} - \cdots] \\ &\quad \times [G_0(-\omega)\delta_{kj} - G_0^2(-\omega)A_{kj} + G_0^3(-\omega)A_{kl}A_{lj} - \cdots] \\ &= G_0(\omega)G_0(-\omega)\delta_{ij} - [G_0^2(\omega)G_0(-\omega) + G_0(\omega)G_0^2(-\omega)]A_{ij} \\ &\quad + [G_0^3(\omega)G_0(-\omega) + G_0^2(\omega)G_0^2(-\omega) + G_0(\omega)G_0^3(-\omega)]A_{ik}A_{kj} - \cdots . \end{aligned} \tag{A5}$$

The quantities in the square brackets can easily be summed. $G_0(-\omega)$ is equal to the complex conjugate of $G_0(\omega)$, $\bar{G}_0(\omega) = G_0(-\omega)$, so

$$G_0^n\bar{G}_0 + G_0^{n-1}\bar{G}_0^2 + \cdots + G_0\bar{G}_0^n = G_0\bar{G}_0\frac{G_0^n - \bar{G}_0^n}{G_0 - \bar{G}_0} = G_0\bar{G}_0\frac{\operatorname{Im}G_0^n}{\operatorname{Im}G_0}. \tag{A6}$$

Then

$$\begin{aligned} g_{ij}(\omega)g_{jk}(-\omega) &= \frac{G_0(\omega)\bar{G}_0(\omega)}{\operatorname{Im}G_0(\omega)}[\operatorname{Im}G_0(\omega)\delta_{ij} \\ &\quad - \operatorname{Im}G_0^2(\omega)A_{ij} + \operatorname{Im}G_0^3(\omega)A_{ik}A_{kj} - \ldots] \\ &= \frac{\gamma_0}{\omega}\operatorname{Im}g_{ij}(\omega). \end{aligned} \tag{A7}$$

Finally (A4) gives the usual form for a fluctuation dissipation theorem (FDT)

$$c_{ij}(\omega) = \frac{2T}{\omega}\operatorname{Im}g_{ij}(\omega). \tag{A8}$$

This can be integrated by using the Kramers–Kronig relations (see [11]) to give

$$c_{ij}(\tau = 0) = Tg_{ij}(\omega = 0). \tag{A9}$$

For output noise given by (29) a similar calculation leads to (30). To see that this fluctuation-response relation is the same as the one derived in [7] one has to use equation (18):

$$C \equiv \left[\frac{1}{N}\operatorname{tr}\mathbf{c}\right]_\xi = T(1 - \lambda G) = \frac{T\alpha G}{1 + G} \tag{A10}$$

where G means $G(\omega = 0)$.

Appendix B. Diagram expansion for input noise

In this appendix the leading corrections to the self-energy will be found. $\gamma_0 = 1$ will be assumed.

There are two families of diagrams in the self-energy, see figure 2. The family where the two ends are connected by a pattern average can be summed exactly; these include the original diagrams from the noiseless case, shown in figure 1. The first of these (first in the second row of figure 2) gives, except for a minus sign,

$$\frac{\alpha}{2\pi}\int \mathrm{d}\omega''\,(\epsilon(\omega-\omega'')\epsilon(\omega''-\omega'))_\gamma G(\omega'') \simeq \frac{\alpha\gamma}{\Delta-\mathrm{i}\omega}2\pi\delta(\omega-\omega'). \tag{B1}$$

Here it was used that $G(t) = 0$ for $t < 0$ (causality), and that $G(t=0) = 1$. It was also assumed that $G(t)$ varies on a time scale much slower than $1/\Delta$, that is the largest eigenvalue of $\mathbf{A}^\epsilon$ has to be much smaller that Δ. To lowest order the input noise corresponds to a weight decay of size $\gamma\alpha$, which implies that the largest eigenvalue is $\gamma\alpha + (1+\sqrt{\alpha})^2$. Therefore the condition is

$$\Delta \gg \gamma\alpha + (1+\sqrt{\alpha})^2. \tag{B2}$$

This condition will also apply to the rest of the calculations here.

By calculating a few more diagrams from this family, as done earlier, it turns out that each noise average just contributes $\gamma/(\Delta-\mathrm{i}\omega)$, so the sum of the whole family can be written as

$$\alpha\left(1-\left[G(\omega)+\frac{\gamma}{\Delta-\mathrm{i}\omega}\right]+\left[G(\omega)+\frac{\gamma}{\Delta-\mathrm{i}\omega}\right]^2-\ldots\right)\delta(\omega-\omega') = \frac{\alpha}{1+G(\omega)+\gamma/(\Delta-\mathrm{i}\omega)}\delta(\omega-\omega'). \tag{B3}$$

This includes terms of higher order than $1/\Delta$, which of course does not hurt.

It is not straightforward to sum the other family of diagrams with a noise average connecting the ends (first row of diagrams in figure 2). Therefore only the lowest order ones are considered. Apart from the first of these, which is calculated in (37), only the two following diagrams contributes to order $1/\Delta$. The first of these gives the same as (B1), and the second gives (except for a minus sign)

$$\frac{\alpha}{(2\pi)^3}\int \mathrm{d}\omega''\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\,(\epsilon(\omega_1)\epsilon(\omega_2))_\gamma(\epsilon(\omega-\omega''-\omega_1)\epsilon(\omega''-\omega'-\omega_2))_\gamma G(\omega'') \simeq \frac{\gamma^2\alpha}{2\Delta-\mathrm{i}\omega}2\pi\delta(\omega-\omega'). \tag{B4}$$

Higher order diagrams can be calculated in the same way, but they turn out to be smaller by at least $1/\Delta$.

Finally the terms can be collected to give the self-energy,

$$\Sigma = \alpha\gamma - \frac{\alpha\gamma}{\Delta-\mathrm{i}\omega} - \frac{\alpha\gamma^2}{2\Delta-\mathrm{i}\omega} + \frac{\alpha}{1+G+\gamma/(\Delta-\mathrm{i}\omega)} \tag{B5}$$

where the $2\pi\delta(\omega - \omega')$ is dropped. Because Δ is large, Σ can be approximated by setting $\omega = 0$ in the new terms,

$$\Sigma = \alpha\gamma\left(1 - \frac{1+\gamma/2}{\Delta}\right) + \frac{\alpha}{1+G+\gamma/\Delta}. \tag{B6}$$

In this approximation the weight decay is just a little smaller than was found to lowest order in Δ. It is actually easy to solve for G. Putting $\lambda = \alpha\gamma[1 - (1+\gamma/2)/\Delta]$ equation (14) reads

$$G^{-1} = \lambda - i\omega + \frac{\alpha/(1+\gamma/\Delta)}{1+G/(1+\gamma/\Delta)}. \tag{B7}$$

If this is compared to (17) we see that it works exactly like having a different (reduced) static variance of the input patterns, $\sigma^2 = 1/[1+\gamma/\Delta]$.

Thus it has been shown that input noise corresponds (to order $1/\Delta$) to having a weight decay of size $\gamma\alpha[1-(1+\gamma/2)/\Delta)]$ and a variance of the input distribution of $1/[1+\gamma/\Delta)]$ instead of 1.

References

[1] Cover T M 1965 *IEEE Trans. Electron. Comput.* **14** 326–34
[2] Amit D, Gutfreund H and Sompolinsky H 1987 *Ann. Phys.* **173** 30–67
[3] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257–70
[4] Hertz J A, Krogh A and Thorbergsson G I 1989 *J. Phys. A: Math. Gen.* **22** 2133–50
[5] Opper M 1989 *Europhys. Lett.* **8** 389–92
[6] Kinzel W and Opper M 1990 *Physics of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)
[7] Der R 1990 *J. Phys. A: Math. Gen.* **23** L763–6
[8] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City: Addison-Wesley)
[9] Kohonen T 1989 *Self-Organization and Associative Memory* (Berlin: Springer)
[10] Levin E, Tishby N and Solla S A 1990 *IEEE Proc.* **78** 1568–74
[11] Landau L D and Lifschitz E M 1980 *Statistical Physics I* (Oxford: Pergamon)
[12] Krogh A and Hertz J A 1991 *Neural Information Processing Systems 3* ed R P Lippmann *et al* (San Mateo: Morgan Kaufmann) pp 897–903
[13] Krogh A and Hertz J A 1991 *J. Phys. A: Math. Gen.* **25** 1135–47