Analog Neural Nets with Gaussian or other Common Noise Distributions cannot Recognize Arbitrary Regular Languages

Wolfgang Maass
Inst. for Theoretical Computer Science,
Technische Universität Graz
Klosterwiesgasse 32/2,
A-8010 Graz, Austria
email: maass@igi.tu-graz.ac.at

Eduardo D. Sontag
Dep. of Mathematics
Rutgers University
New Brunswick, NJ 08903, USA
email: sontag@hilbert.rutgers.edu

Abstract

We consider recurrent analog neural nets where the output of each gate is subject to Gaussian noise, or any other common noise distribution that is nonzero on a sufficiently large part of the state space. We show that many regular languages cannot be recognized by networks of this type, and we give a precise characterization of those languages which can be recognized. This result implies severe constraints on possibilities for constructing recurrent analog neural nets that are robust against realistic types of analog noise. On the other hand we present a method for constructing feedforward analog neural nets that are robust with regard to analog noise of this type.

1 Introduction

A fairly large literature (see [Omlin, Giles, 1996] and the references therein) is devoted to the construction of analog neural nets that recognize regular languages. Any physical realization of the analog computational units of an analog neural net in technological or biological systems is bound to encounter some form of "imprecision" or analog noise at its analog computational units. We show in this article that this effect has serious consequences for the capability of analog neural nets with regard to language recognition. We show that any analog neural net whose analog computational units are subject to Gaussian or other common noise distributions cannot recognize arbitrary regular languages. For example, such analog neural net cannot recognize the regular language $\{w \in \{0,1\}^* | w \text{ begins with } 0\}$.

A precise characterization of those regular languages which can be recognized by such analog neural nets is given in Theorem 1.1. In section 3 we introduce a simple technique for making feedforward neural nets robust with regard to the here considered types of analog noise. This method is employed to prove the positive part of Theorem 1.1. The main difficulty in proving Theorem 1.1 is its negative part, for which adequate theoretical tools are introduced in section 2. The proof of this negative part of Theorem 1.1 holds for quite general stochastic analog computational systems. However for the sake of simplicity we will tailor our description towards the special case of noisy neural networks.

Before we can give the exact statement of Theorem 1.1 and discuss related preceding work we have to give a precise definition of computations in noisy neural networks. From the conceptual point of view this definition is basically the same as for computations in noisy boolean circuits (see [Pippenger, 1985] and [Pippenger, 1990]). However it is technically more involved since we have to deal here with an infinite state space. Recognition of a language $L \subseteq U^*$ by a noisy analog computational system M with discrete time is defined essentially as in [Maass, Orponen, 1997]. The set of possible internal states of M is assumed to be some (Borel) measurable set $\Omega \subseteq \mathbb{R}^n$, for some integer n (which is called the "number of neurons" or the "dimension"). A typical choice is $\Omega = [-1,1]^n$. The input set is the alphabet U. We assume given an auxiliary mapping

$$f : \Omega \times U \to \widehat{\Omega}$$

which describes the transitions in the absence of noise (and saturation effects), where $\widehat{\Omega} \subseteq \mathbb{R}^n$ is an intermediate set which is (Borel) measurable, and $f(\cdot, u)$ is supposed to be (Borel) measurable for each fixed $u \in U$. The system description is completed by specifying a stochastic kernel[1] $Z(\cdot,\cdot)$ on $\widehat{\Omega} \times \Omega$. We interpret $Z(y, A)$ as the probability that a vector y can be corrupted by noise (and possibly truncated in values) into a state in the set A. The probability of transitions from a state $x \in \Omega$ to a set $A \subseteq \Omega$, if the current input value is u, is defined, in terms of this data, as:

$$\mathbb{K}_u(x, A) := Z(f(x,u), A)\,.$$

This is itself a stochastic kernel, for each given u.

More specifically for this paper, we assume that the noise kernel $Z(y, A)$ is given in terms of an additive "noise" or "error" $\mathbb{R}^n$-valued random variable V with density $\phi(\cdot)$, and a fixed (Borel) measurable "saturation" function

$$\sigma : \mathbb{R}^n \to \Omega\,,$$

[1] That is to say, $Z(y, A)$ is defined for each $y \in \widehat{\Omega}$ and each (Borel) measurable subset $A \subseteq \Omega$, $Z(y, \cdot)$ is a probability distribution for each y, and $Z(\cdot, A)$ is a measurable function for each A.

as follows. For any $y \in \widehat{\Omega}$ and any $A \subseteq \Omega$, let A_y denote the set

$$\sigma^{-1}(A) - \{y\} := \{x - y \mid \sigma(x) \in A\}\,.$$

(Also, generally for any $A, B \subseteq \mathbb{R}^n$, let $A - B$ denote the set of all possible differences of elements of A and B.) Then the kernel Z is defined as:

$$Z(y, A) := \mathrm{Prob}_\phi\,[\sigma(y + V) \in A] = \int_{A_y} \phi(v)\,dv\,.$$

The main assumption throughout this article is that the noise has a wide support. To be precise: *There exist a subset Ω_0 of Ω, and some constant $c_0 > 0$, such that the following two properties hold:*

$$\widehat{\Omega}_0 := \sigma^{-1}(\Omega_0) \text{ has finite and nonzero Lebesgue measure } m_0 = \lambda\left(\widehat{\Omega}_0\right) \tag{1}$$

and

$$\phi(v) \geq c_0 \text{ for all } v \in Q := \widehat{\Omega}_0 - \widehat{\Omega}\,. \tag{2}$$

A special (but typical) case of this situation is that in which $\Omega = [-1,1]^n$ and σ is the "standard saturation" in which each coordinate is projected onto the interval $[-1,1]$. That is, for real numbers z, we let $\mathrm{sat}\,(z) = \mathrm{sign}\,(z)$ if $|z| > 1$ and $\mathrm{sat}\,(z) = z$ if $|z| \leq 1$, and for vectors $y = (y_1, \ldots, y_n)' \in \mathbb{R}^n$ we let $\sigma(y) := (\mathrm{sat}\,(y_1), \ldots, \mathrm{sat}\,(y_n))$. In that case, provided that the density ϕ of V is continuous and satisfies

$$\phi(v) \neq 0 \text{ for all } v \in \Omega - \widehat{\Omega}\,, \tag{3}$$

and assuming that $\widehat{\Omega}$ is compact, both assumptions (2) and (1) are satisfied. Indeed, we may pick as our set Ω_0 any subset $\Omega_0 \subseteq (-1,1)^n$ with nonzero Lebesgue measure. Then, $\sigma^{-1}(\Omega_0) = \Omega_0$, and, since ϕ is continuous and everywhere nonzero on the compact set $\Omega - \widehat{\Omega} \supset \Omega_0 - \widehat{\Omega} = Q$, there is a constant $c_0 > 0$ as desired. Obviously, condition (3) is satisfied by the probability density function of the Gaussian distribution (and many other common distributions that are used to model analog noise) since these density functions satisfy $\phi(v) \neq 0$ for all $v \in \mathbb{R}^n$.

The main example of interest is that of (first order or high order) neural networks. In the case of first order neural networks one takes a bounded (usually, two-element) $U \subseteq \mathbb{R}$, $\Omega = [-1,1]^n$, and

$$f : [-1,1]^n \times U \to \widehat{\Omega} \subseteq \mathbb{R}^n : (x, u) \mapsto Wx + h + uc\,, \tag{4}$$

where $W \in \mathbb{R}^{n \times n}$ and $c, h \in \mathbb{R}^n$ represent weight matrix and vectors, and $\widehat{\Omega}$ is any compact subset which contains the image of f. The complete noisy neural network model is thus described by transitions

$$x_{t+1} = \sigma(Wx_t + h + u_t c + V_t)\,,$$

where $V_1, V_2, \ldots$ is a sequence of independent random n-vectors, all distributed identically to V; for example, $V_1, V_2, \ldots$ might be an i.i.d. Gaussian process.

A variation of this example is that in which the noise affects the activation *after* the desired transition, that is, the new state is

$$x_{t+1} = \sigma(Wx_t + h + u_t c) + V_t\,,$$

again with each coordinate clipped to the interval $[-1, 1]$. This can be modeled as

$$x_{t+1} = \sigma(\sigma(Wx_t + h + u_t c) + V_t)\,,$$

and becomes a special case of our setup if we simply let

$$f(x, u) = \sigma(Wx_t + h + u_t c)\,.$$

For each (signed, Borel) measure μ on Ω, and each $u \in U$, we let $\mathbb{K}_u \mu$ be the (signed, Borel) measure defined on Ω by $(\mathbb{K}_u \mu)(A) := \int K_u(x, A) d\mu(x)$. Note that $\mathbb{K}_u \mu$ is a probability measure whenever μ is. For any sequence of inputs $w = u_1, \ldots, u_r$, we consider the composition of the evolution operators $\mathbb{K}_{u_i}$:

$$\mathbb{K}_w = \mathbb{K}_{u_r} \circ \mathbb{K}_{u_{r-1}} \circ \ldots \circ \mathbb{K}_{u_1}\ . \tag{5}$$

If the probability distribution of states at any given instant is given by the measure μ, then the distribution of states after a single computation step on input $u \in U$ is given by $\mathbb{K}_u \mu$, and after r computation steps on inputs $w = u_1, \ldots, u_r$, the new distribution is $\mathbb{K}_w \mu$, where we are using the notation (5). In particular, if the system starts at a particular initial state ξ, then the distribution of states after r computation steps on w is $\mathbb{K}_w \delta_\xi$, where δ_ξ is the probability measure concentrated on $\{\xi\}$. That is to say, for each measurable subset $F \subseteq \Omega$

$$\text{Prob}\,[x_{r+1} \in F \,|\, x_1 = \xi,\ \text{input } = w] \;=\; (\mathbb{K}_w \delta_\xi)(F)\,.$$

We fix an initial state $\xi \in \Omega$, a set of "accepting" or "final" states F, and a "reliability" level $\varepsilon > 0$, and say that $M = (M, \xi, F, \varepsilon)$ *recognizes the subset* $L \subseteq U^*$ if for all $w \in U^*$:

$$\begin{aligned} w \in L &\iff (\mathbb{K}_w \delta_\xi)(F) \geq \frac{1}{2} + \varepsilon \\ w \notin L &\iff (\mathbb{K}_w \delta_\xi)(F) \leq \frac{1}{2} - \varepsilon\,. \end{aligned}$$

In general a neural network that simulates a DFA will carry out not just one, but a fixed number k of computation steps (=state transitions) of the form $x' = \text{sat}(Wx + h + uc) + V$ for each input symbol $u \in U$ that it reads (see the constructions described in [Omlin, Giles, 1996], and in section 3 of this article). This can easily be reflected in our model by formally replacing any input sequence $w = u_1, u_2, \ldots, u_r$ from U^* by a padded sequence $\tilde{w} = u_1, b^{k-1}, u_2, b^{k-1}, \ldots, u_r, b^{k-1}$ from $(U \cup \{b\})^*$, where b is a blank symbol not in U, and b^{k-1} denotes a sequence of $k - 1$ copies of b (for some arbitrarily fixed $k \geq 1$). Then one defines

$$\begin{aligned} w \in L &\iff (\mathbb{K}_{\tilde{w}} \delta_\xi)(F) \geq \frac{1}{2} + \varepsilon \\ w \notin L &\iff (\mathbb{K}_{\tilde{w}} \delta_\xi)(F) \leq \frac{1}{2} - \varepsilon\,. \end{aligned}$$

This completes our definition of language recognition by a noisy analog computational system M with discrete time. This definition agrees with that given in [Maass, Orponen, 1997].

The main result of this article is the following:

Theorem 1.1 *Assume that U is some arbitrary finite alphabet. A language $L \subseteq U^*$ can be recognized by a noisy analog computational system M of the previously specified type if and only if $L = E_1 \bigcup U^* E_2$ for two finite subsets E_1 and E_2 of U^*.*

As an illustration of the statement of Theorem 1.1 we would like to point out that it implies for example that the regular language $L = \{w \in \{0,1\}^* | w \text{ begins with } 0\}$ cannot be recognized by a noisy analog computational system, but the regular language $L = \{w \in \{0,1\}^* | w \text{ ends with } 0\}$ can be recognized by such system. The *proof* of Theorem 1.1 follows immediately from Corollary 2.2 and Corollary 3.3.

A corresponding version of Theorem 1.1 for *discrete* computational systems was previously shown in [Rabin, 1963]. More precisely, Rabin had shown that probabilistic automata with strictly positive matrices can recognize exactly the same class of languages L that occur in our Theorem 1.1. Rabin referred to these languages as *definite languages.* Language recognition by *analog* computational systems with *analog* noise has previously been investigated in [Casey, 1996] for the special case of bounded noise and perfect reliability (i.e. $\int_{\|v\| \leq \eta} \phi(v) dv = 1$ for some small $\eta > 0$ and $\varepsilon = 1/2$ in our terminology), and in [Maass, Orponen, 1997] for the general case. It was shown in [Maass, Orponen, 1997] that any such system can only recognize regular languages. Furthermore it was shown there that if $\int_{\|v\| \leq \eta} \phi(v) dv = 1$ for some small $\eta > 0$ then all regular languages *can* be recognized by such systems. In the present paper we focus on the complementary case where the condition "$\int_{\|v\| \leq \eta} \phi(v) dv = 1$ for some small $\eta > 0$" is not satisfied, i.e. analog noise may move states over larger distances in the state space. We show that even if the probability of such event is arbitrarily small, the neural net will no longer be able to recognize arbitrary regular languages.

2 A Constraint on Language Recognition

We prove in this section the following result for arbitrary noisy computational systems M as defined in section 1:

Theorem 2.1 *Assume that U is some arbitrary finite alphabet. If a language $L \subseteq U^*$ is recognized by M, then there are subsets E_1 and E_2 of $U^{\leq r}$, for some integer r, such that $L = E_1 \bigcup U^* E_2$. In other words: whether a string $w \in U^*$ belongs to the language L can be decided by just inspecting the first r and the last r symbols in w.*

Corollary 2.2 *Assume that U is a finite alphabet. If L is recognized by M, then there are finite subsets E_1 and E_2 of U^* such that $L = E_1 \bigcup U^* E_2$.* ∎

Remark 2.3 *The result is also true in various cases when the noise random variable is not be necessarily independent of the new state $f(x, u)$. The proof depends only on the fact that the kernels K_u satisfy the Doeblin condition with a uniform constant (see Lemma 2.5 in the next section).*

2.1 A General Fact about Stochastic Kernels

Let $(S, \mathcal{S})$ be a measure space, and let K be a stochastic kernel. As in the special case of the K_u's above, for each (signed) measure μ on $(S, \mathcal{S})$, we let $\mathbb{K}\mu$ be the (signed) measure defined on $\mathcal{S}$ by $(\mathbb{K}\mu)(A) := \int K(x, A) d\mu(x)$. Observe that $\mathbb{K}\mu$ is a probability measure whenever μ is. Let $c > 0$ be arbitrary. We say that K satisfies *Doeblin's condition (with constant c)* if there is some probability measure ρ on $(S, \mathcal{S})$ so that

$$K(x, A) \geq c\rho(A) \quad \text{for all } x \in S, A \in \mathcal{S}\,. \tag{6}$$

(Necessarily $c \leq 1$, as is seen by considering the special case $A = S$.) This condition is due to [Doeblin, 1937].

We denote by $\|\mu\|$ the *total variation* of the (signed) measure μ. Recall that $\|\mu\|$ is defined as follows. One may decompose S into a disjoint union of two sets A and B, in such a manner that μ is nonnegative on A and nonpositive on B. Letting the restrictions of μ to A and B be "μ_+" and "$-\mu_-$" respectively (and zero on B and A respectively), we may decompose μ as a difference of nonnegative measures with disjoint supports, $\mu = \mu_+ - \mu_-$. Then, $\|\mu\| = \mu_+(A) + \mu_-(B)$.

The following Lemma is a "folk" fact ([Papinicolaou, 1978]), but we have not been able to find a proof in the literature; thus, we provide a self-contained proof.

Lemma 2.4 *Assume that K satisfies Doeblin's condition with constant c. Let μ be any (signed) measure such that $\mu(S) = 0$. Then,*

$$\|\mathbb{K}\mu\| \leq (1 - c)\, \|\mu\|\,. \tag{7}$$

Proof. In terms of the above decomposition of μ, $\mu(S) = 0$ means that $\mu_+(A) = \mu_-(B)$. We denote $q := \mu_+(A) = \mu_-(B)$. Thus, $\|\mu\| = 2q$. If $q = 0$ then $\mu \equiv 0$, and so also $\mathbb{K}\mu \equiv 0$ and there is nothing to prove. So from now on we assume $q \neq 0$. Let $\nu_1 := K\mu_+$, $\nu_2 := \mathbb{K}\mu_-$, and $\nu := \mathbb{K}\mu$. Then, $\nu = \nu_1 - \nu_2$. Since $(1/q)\mu_+$ and $(1/q)\mu_-$ are probability measures, $(1/q)\nu_1$ and $(1/q)\nu_2$ are probability measures as well. That is,

$$\nu_1(S) = \nu_2(S) = q\,. \tag{8}$$

We now decompose S into two disjoint measurable sets C and D, in such a fashion that $\nu_1 - \nu_2$ is nonnegative on C and nonpositive on D. So

$$\begin{aligned}\|\nu\| = (\nu_1 - \nu_2)(C) + (\nu_2 - \nu_1)(D) &= \nu_1(C) - \nu_1(D) + \nu_2(D) - \nu_2(C) \\ &= 2q - 2\nu_1(D) - 2\nu_2(C)\,,\end{aligned} \tag{9}$$

where we used that $\nu_1(D)+\nu_1(C)=q$ and similarly for ν_2. By Doeblin's condition,

$$\nu_1(D)=\int K(x,D)d\mu_+(x)\geq c\rho(D)\int d\mu_+(x)=c\rho(D)\mu_+(A)=cq\rho(D).$$

Similarly, $\nu_2(C)\geq cq\rho(C)$. Therefore, $\nu_1(D)+\nu_2(C)\geq cq$ (recall that $\rho(C)+\rho(D)=1$, because ρ is a probability measure). Substituting this last inequality into Equation (9), we conclude that $\|\nu\|\leq 2q-2cq=(1-c)2q=(1-c)\|\mu\|$, as desired. ∎

2.2 Proof of Theorem 2.1

The main technical observation regarding the measure K_u defined in section 1 is as follows.

Lemma 2.5 *There is a constant $c>0$ such that K_u satisfies Doeblin's condition with constant c, for every $u\in U$.*

Proof. Let Ω_0, c_0, and $0<m_0<1$ be as in assumptions (2) and (1), and introduce the following (Borel) probability measure on Ω_0:

$$\lambda_0(A) := \frac{1}{m_0}\lambda\left(\sigma^{-1}(A)\right).$$

Pick any measurable $A\subseteq\Omega_0$ and any $y\in\widehat{\Omega}$. Then,

$$\begin{aligned} Z(y,A) &= \operatorname{Prob}[\sigma(y+V)\in A] = \operatorname{Prob}[y+V\in\sigma^{-1}(A)] \\ &= \int_{A_y}\phi(v)\,dv \geq c_0\lambda(A_y) = c_0\lambda\left(\sigma^{-1}(A)\right) = c_0m_0\lambda_0(A), \end{aligned}$$

where $A_y:=\sigma^{-1}(A)-\{y\}\subseteq Q$. We conclude that $Z(y,A)\geq c\lambda_0(A)$ for all y,A, where $c=c_0m_0$. Finally, we extend the measure λ_0 to all of Ω by assigning zero measure to the complement of Ω_0, that is, $\rho(A):=\lambda_0(A\bigcap\Omega_0)$ for all measurable subsets A of Ω. Pick $u\in U$; we will show that K_u satisfies Doeblin's condition with the above constant c (and using ρ as the "comparison" measure in the definition). Consider any $x\in\Omega$ and measurable $A\subseteq\Omega$. Then,

$$K_u(x,A)=Z(f(x,u),A)\geq Z(f(x,u),A\bigcap\Omega_0)\geq c\lambda_0(A\bigcap\Omega_0)=c\rho(A),$$

as required. ∎

For every two probability measures μ_1,μ_2 on Ω, applying Lemma 2.4 to $\mu:=\mu_1-\mu_2$, we know that $\|\mathbb{K}_u\mu_1-\mathbb{K}_u\mu_2\|\leq(1-c)\|\mu_1-\mu_2\|$ for each $u\in U$. Recursively, then, we conclude:

$$\|\mathbb{K}_w\mu_1-\mathbb{K}_w\mu_2\|\leq(1-c)^r\|\mu_1-\mu_2\|\leq 2(1-c)^r \tag{10}$$

for all words w of length $\geq r$.

Now pick any integer r such that $(1-c)^r<2\varepsilon$. From Equation (10), we have that

$$\|\mathbb{K}_w\mu_1-\mathbb{K}_w\mu_2\|<4\varepsilon$$

for all w of length $\geq r$ and any two probability measures μ_1, μ_2. In particular, this means that, for each measurable set A,

$$|(\mathbb{K}_w \mu_1)(A) - (\mathbb{K}_w \mu_2)(A)| < 2\varepsilon \tag{11}$$

for all such w. (Because, for any two probability measures ν_1 and ν_2, and any measurable set A, $2\,|\nu_1(A) - \nu_2(A)| \leq \|\nu_1 - \nu_2\|$.)

We denote by $w_1 w_2$ the concatenation of sequences $w_1, w_2 \in U^*$.

Lemma 2.6 *Pick any $v \in U^*$ and $w \in U^r$. Then*

$$w \in L \iff vw \in L \,.$$

Proof. Assume that $w \in L$, that is, $(\mathbb{K}_w \delta_\xi)(F) \geq \frac{1}{2} + \varepsilon$. Applying inequality (11) to the measures $\mu_1 := \delta_\xi$ and $\mu_2 := \mathbb{K}_v \delta_\xi$ and $A = F$, we have that $|(\mathbb{K}_w \delta_\xi)(F) - (\mathbb{K}_{vw} \delta_\xi)(F)| < 2\varepsilon$, and this implies that $(\mathbb{K}_{vw} \delta_\xi)(F) > \frac{1}{2} - \varepsilon$, i.e., $vw \in L$. (Since $\frac{1}{2} - \varepsilon < (\mathbb{K}_{vw} \delta_\xi)(F) < \frac{1}{2} + \varepsilon$ is ruled out.) If $w \notin L$, the argument is similar. ∎

We have proved that

$$L \bigcap (U^* U^r) \;=\; U^* (L \bigcap U^r)\,.$$

So,

$$L \;=\; \left(L \bigcap U^{\leq r}\right) \bigcup \left(L \bigcap U^* U^r\right) \;=\; E_1 \bigcup U^* E_2$$

where $E_1 := L \bigcap U^{\leq r}$ and $E_2 := L \bigcap U^r$ are both included in $U^{\leq r}$. This completes the proof of Theorem 2.1.

3 Construction of Noise Robust Analog Neural Nets

In this section we exhibit a method for making feedforward analog neural nets robust with regard to arbitrary analog noise of the type considered in the preceding sections. This method can be used to prove in Corollary 3.3 the missing positive part of the claim of the main result (Theorem 1.1) of this article.

Theorem 3.1 *Let $\mathcal{C}$ be any (noiseless) feedforward threshold circuit, and let $\sigma : \mathbb{R} \to [-1, 1]$ be some arbitrary function with $\sigma(u) \to 1$ for $u \to \infty$ and $\sigma(u) \to -1$ for $u \to -\infty$. Furthermore assume that $\delta, \rho \in (0, 1)$ are some arbitrary given parameters. Then one can transform the noiseless threshold circuit $\mathcal{C}$ into an analog neural net $\mathcal{N}_\mathcal{C}$ with the same number of gates, whose gates employ the given function σ as activation function, so that for any analog noise of the type considered in section 1 and any circuit input $\underline{x} \in \{-1, 1\}^m$ the output of $\mathcal{N}_\mathcal{C}$ differs with probability $\geq 1 - \delta$ by at most ρ from the output of $\mathcal{C}$.*

Proof. We can assume that for any threshold gate g in $\mathcal{C}$ and any input $\underline{y} \in \{-1,1\}^l$ to gate g the weighted sum of inputs to gate g has distance ≥ 1 from the threshold of g. This follows from the fact that without loss of generality the weights of g can be assumed to be even integers. Let n be the number of gates in $\mathcal{C}$ and let V be an arbitrary noise vector as described in section 1. In fact, V may be any $\mathbb{R}^n$-valued random variable with some density function $\phi(\cdot)$. Let k be the maximal fan-in of a gate in $\mathcal{C}$, and let w be the maximal absolute value of a weight in $\mathcal{C}$.

We choose $R > 0$ so large that

$$\int_{|v_i| \geq R} \phi(v)\, dv \quad \leq \quad \frac{\delta}{2n} \;\; \text{for} \;\; i = 1, \ldots, n \; .$$

Furthermore we choose $u_0 > 0$ so large that $\sigma(u) \geq 1 - \rho/(wk)$ for $u \geq u_0$ and $\sigma(u) \leq -1 + \rho/(wk)$ for $u \leq -u_0$. Finally we choose a factor $\gamma > 0$ so large that $\gamma(1-\rho) - R \geq u_0$. Let $\mathcal{N}_{\mathcal{C}}$ be the analog neural net that results from $\mathcal{C}$ through multiplication of all weights and thresholds with γ and through replacement of the Heaviside activation functions of the gates in $\mathcal{C}$ by the given activation function σ.

We show that for any circuit input $\underline{x} \in \{-1,1\}^m$ the output of $\mathcal{N}_{\mathcal{C}}$ differs with probability $\geq 1-\rho$ by at most ρ from the output of $\mathcal{C}$, in spite of analog noise V with density $\phi(\cdot)$ in the analog neural net $\mathcal{N}_{\mathcal{C}}$. By choice of R the probability that any of the n components of the noise vector V has an absolute value larger than R is at most $\delta/2$. On the other hand one can easily prove by induction on the depth of a gate g in $\mathcal{C}$ that if all components of V have absolute values $\leq R$ then for any circuit input $\underline{x} \in \{-1,1\}^m$ the output of the analog gate $\tilde{g}$ in $\mathcal{N}_{\mathcal{C}}$ that corresponds to g differs by at most $\rho/(wk)$ from the output of the gate g in $\mathcal{C}$. The induction hypothesis implies that the inputs of $\tilde{g}$ differ by at most $\rho/(wk)$ from the corresponding inputs of g. Therefore the difference of the weighted sum and the threshold at $\tilde{g}$ has a value $\geq \gamma \cdot (1-\rho)$ if the corresponding difference at g has a value ≥ 1, and a value $\leq -\gamma \cdot (1-\rho)$ if the corresponding difference at g has a value ≤ -1. Since the component of the noise vector V that defines the analog noise in gate $\tilde{g}$ has by assumption an absolute value $\leq R$, the output of $\tilde{g}$ is $\geq 1 - \rho/(wk)$ in the former case and $\leq -1 + \rho/(wk)$ in the latter case. Hence it deviates by at most $\rho/(wk)$ from the output of gate g in $\mathcal{C}$. ∎

Remark 3.2

(a) *Any boolean circuit $\mathcal{C}$ with gates for OR, AND, NOT or NAND is a special case of a threshold circuit. Hence one can apply Theorem 3.1 to such circuit.*

(b) *It is obvious from the proof that Theorem 3.1 also holds for circuits with recurrencies, provided that there is a fixed bound T for the computation time of such circuit.*

(c) *It is more difficult to make analog neural nets robust against another type of noise where at each sigmoidal gate the noise is applied* after *the activation σ. With the notation from section 1 of this article this other model can be described by*

$$x_{t+1} = \; sat \; (\sigma(Wx_t + h + u_t c) + V_t) \; .$$

For this noise model it is apparently not possible to prove positive results like Theorem 3.1 without further assumptions about the density function $\phi(v)$ of the noise vector V. However if one assumes that for any i the integral $\int_{|v_i| \geq \rho/(2wk)} \phi(v)dv$ can be bounded by a sufficiently small constant (which can be chosen independently of the size of the given circuit), then one can combine the argument from the proof of Theorem 3.1 with standard methods for constructing boolean circuits that are robust with regard to common models for digital *noise (see for example [Pippenger, 1985], [Pippenger, 1989], [Pippenger, 1990]). In this case one chooses u_0 so that $\sigma(u) \geq 1 - \rho/(2wk)$ for $u \geq u_0$ and $\sigma(u) \leq 1 + \rho/(2wk)$ for $u \leq -u_0$, and multiplies all weights and thresholds of the given threshold circuit with a constant γ so that $\gamma \cdot (1 - \rho) \geq u_0$. One handles the rare occurrences of components $\tilde{V}$ of the noise vector V that satisfy $|\tilde{V}| > \rho/(2wk)$ like the rare occurrences of gate failures in a digital circuit. In this way one can simulate any given feedforward threshold circuit by an analog neural net that is robust with respect to this different model for analog noise.*

The following Corollary provides the proof of the positive part of our main result Theorem 1.1.

Corollary 3.3 *Assume that U is some arbitrary finite alphabet, and language $L \subseteq U^*$ is of the form $L = E_1 \bigcup U^* E_2$ for two arbitrary finite subsets E_1 and E_2 of U^*. Then the language L can be recognized by a noisy analog neural net $\mathcal{N}$ with any desired reliability $\varepsilon \in (0, \frac{1}{2})$, in spite of arbitrary analog noise of the type considered in section 1.*

Proof. On the basis of Theorem 3.1 the proof of this Corollary is rather straightforward. We first construct a feedforward threshold circuit $\mathcal{C}$ for recognizing L, that receives each input symbol from U in the form of a bitstring $u \in \{0,1\}^l$ (for some fixed $l \geq \log_2 |U|$), that is encoded as the binary states of l input units of the boolean circuit $\mathcal{C}$. Via a tapped delay line of fixed length d (which can easily be implemented in a feedforward threshold circuit by d layers, each consisting of l gates that compute the idendity function of a single binary input from the preceding layer) one can achieve that the feedforward circuit $\mathcal{C}$ computes *any* given boolean function of the last d sequences from $\{0,1\}^l$ that were presented to the circuit. On the other hand for any language of the form $L = E_1 \cup U^* E_2$ with E_1, E_2 finite there exists some $d \in \mathbb{N}$ such that for each $w \in U^*$ one can decide whether $w \in L$ by just inspecting the last d characters of w. Therefore a feedforward threshold circuit $\mathcal{C}$ with a tapped delay line of the type described above can decide whether $w \in L$.

We apply Theorem 3.1 to this circuit $\mathcal{C}$ for $\delta = \rho = \min(\frac{1}{2} - \varepsilon,\ \frac{1}{4})$. We define the set F of accepting states for the resulting analog neural net $\mathcal{N}_\mathcal{C}$ as the set of those states where the computation is completed and the output gate of $\mathcal{N}_\mathcal{C}$ assumes a value $\geq 3/4$. Then according to Theorem 3.1 the analog neural net $\mathcal{N}_\mathcal{C}$ recognizes L with reliability ε. To be formally precise, one has to apply Theorem 3.1 to a threshold circuit $\mathcal{C}$ that receives its input not in a single batch, but through a sequence of d batches. The proof of Theorem 3.1 readily extends to this case.

Note that according to Theorem 3.1 we may employ as activation functions for the gates of $\mathcal{N}_\mathcal{C}$ arbitrary functions $\sigma : \mathbb{R} \to [-1, 1]$ that satisfy $\sigma(u) \to 1$ for $u \to \infty$ and $\sigma(u) \to -1$ for $u \to -\infty$. ∎

4 Conclusions

We have proven a perhaps somewhat surprising result about the computational power of noisy analog neural nets: analog neural nets with Gaussian or other common noise distributions that are nonzero on a large set cannot accept arbitrary regular languages, even if the mean of the noise distribution is 0, its variance is chosen arbitrarily small, and the reliability $\varepsilon > 0$ of the network is allowed to be arbitrarily small. For example, they cannot accept the regular language $\{w \in \{0,1\}^* | w \text{ begins with } 0\}$. This shows that there is a severe limitation for making recurrent analog neural nets robust against analog noise. The proof of this result introduces new mathematical arguments into the investigation of neural computation, which can also be applied to other stochastic analog computational systems.

Furthermore we have given a precise characterization of those regular languages that can be recognized with reliability $\varepsilon > 0$ by recurrent analog neural nets of this type.

Finally we have presented a method for constructing feedforward analog neural nets that are robust with regard to any of those types of analog noise which are considered in this paper.

5 Acknowledgement

The authors wish to thank Dan Ocone, from Rutgers, for pointing out Doeblin's condition, which resulted in a considerable simplification of their original proof. Also, they gratefully acknowledge an anonymous referee for many useful suggestions regarding presentation.

References

[Casey, 1996] Casey, M., "The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction", *Neural Computation 8*, 1135–1178, 1996.

[Doeblin, 1937] Doeblin, W., "Sur le propriétés asymtotiques de mouvement régis par certain types de chaînes simples", *Bull. Math. Soc. Roumaine Sci. 39(1)*: 57–115; (2) 3–61, 1937.

[Maass, Orponen, 1997] Maass, W., and Orponen, P. "On the effect of analog noise on discrete-time analog computations", *Advances in Neural Information Processing*

Systems 9, 1997, 218–224; journal version: Neural Computation 10, 1998. detailed version see http://www.math.jyu.fi/~orponen/papers/noisyac.ps .

[Omlin, Giles, 1996] Omlin, C. W., Giles, C. L. "Constructing deterministic finite-state automata in recurrent neural networks", *J. Assoc. Comput. Mach. 43* (1996), 937–972.

[Papinicolaou, 1978] Papinicolaou, G., "Asymptotic Analysis of Stochastic Equations", in *Studies in Probability Theory, MAA Studies in Mathematics*, vol. 18, 111–179, edited by M. Rosenblatt, Math. Assoc. of America, 1978.

[Pippenger, 1985] Pippenger, N., "On networks of noisy gates", *IEEE Sympos. on Foundations of Computer Science*, vol. 26, IEEE Press, New York, 30–38, 1985.

[Pippenger, 1989] Pippenger, N., "Invariance of complexity measures for networks with unreliable gates", *J. of the ACM*, vol. 36, 531–539, 1989.

[Pippenger, 1990] Pippenger, N., "Developments in 'The Synthesis of Reliable Organisms from Unreliable Components' ", *Proc. of Symposia in Pure Mathematics*, vol. 50, 311–324, 1990.

[Rabin, 1963] Rabin, M., "Probabilistic automata", *Information and Control*, vol. 6, 230–245, 1963.