LETTER Communicated by Eduardo Sontag

On the Effect of Analog Noise in Discrete-Time Analog Computations

Wolfgang Maass
Institute for Theoretical Computer Science, Technische Universität Graz, Graz, Austria

Pekka Orponen
Department of Mathematics, University of Jyväskylä, Jyväskylä, Finland

We introduce a model for analog computation with discrete time in the presence of analog noise that is flexible enough to cover the most important concrete cases, such as noisy analog neural nets and networks of spiking neurons. This model subsumes the classical model for digital computation in the presence of noise. We show that the presence of arbitrarily small amounts of analog noise reduces the power of analog computational models to that of finite automata, and we also prove a new type of upper bound for the VC-dimension of computational models with analog noise.

1 Introduction

Analog noise is a serious issue in practical analog computation. However, there exists no formal model for reliable computations by noisy analog systems that allows this issue to be addressed in an adequate manner. We propose and investigate such model in this article.

The investigation of noise-tolerant digital computations in the presence of stochastic failures of gates or wires was initiated by von Neumann (1956). We refer to Cowan (1966), Pippenger (1989), and Gál (1991) for a small sample of the numerous results that have been achieved in this direction. In all these articles, one considers computations that produce a correct output not with perfect reliability but with probability $\geq \frac{1}{2} + \rho$ (for some parameter $\rho \in (0, \frac{1}{2}]$). The same framework (with stochastic failures of gates or wires) has been applied to analog neural nets in Siegelmann (1994).

The approaches noted are insufficient for the investigation of noise in analog computations, because one has to be concerned not only with occasional total failures of gates or wires, but also with imprecision—omnipresent smaller (and occasionally larger) perturbations of analog outputs of internal computational units. These perturbations may, for example, be given by gaussian distributions. Therefore, we introduce and investigate in this article a notion of noise-robust computation by noisy analog systems where we assume that the values of intermediate analog values are moved according to some quite arbitrary probability distribution. We consider, as in the traditional framework for noisy digital computations, arbitrary computations whose output is correct with some given probability $\geq \frac{1}{2} + \rho$ (for $\rho \in (0, \frac{1}{2}]$). We restrict our attention to analog computations with digital output. Since we impose no restriction (such as continuity) on the type of operations that can be performed by computational units in an analog computational system, an output unit of such a system can convert an analog value into a binary output via thresholding.

We show in theorem 1 that any language recognized by such noisy analog computational system is regular. Our model and the theorems are somewhat related to the analysis of probabilistic finite automata in Rabin (1963), although in Rabin's case the finiteness of the state-space simplifies the setup considerably. Continuous-space noise models similar to ours have been used in general studies of the stability of dynamical systems affected by random perturbations (e.g., Kifer, 1988), but our work is to our knowledge the first to consider the computational aspects of systems of this type. More specific hardware-oriented models for analog noise in analog neural nets have been discussed in Phatak and Koren (1995).

Another related work is Casey (1996), which addresses the special case of analog computations on recurrent neural nets, where the analog noise can move an internal state at most over some bounded distance η, and the digital output is required to be perfectly reliable ($\rho = 1/2$ in the present notation). Casey's corollary 3.1 states a special case of our theorem 1 for the model considered in that article. Casey's proof of corollary 3.1 is incorrect.[1] A correct proof is contained as a special case in the proof of Theorem 1 in section 3 of this article.[2] Apart from corollary 3.1 there is no further overlap between Casey (1996) and this article.

There are relatively few examples of nontrivial computations on common digital or analog computational models that can achieve perfect reliability of the output in spite of noisy internal components. Most constructions of noise-robust computational models rely on the replication of noisy computational units (see von Neumann, 1956; Cowan, 1966). The idea of this method is that the average of the outputs of k identical noisy networks (with stochastically independent noise processes) is more reliable than the output

[1] Corollary 3.1 is derived as a corollary of theorem 1 in Casey (1996), whose proof relies on the assumption that the recognized language is regular. The proof given for corollary 3.1 is the following: "The proof of a corollary is simply to notice that by the compactness of the phase space it can contain only a finite number of disjoint sets with nonempty interior." The following counterexample shows that this argument is wrong: The intervals $[1/(2i+1), 1/2i]$ for $i = 1, 2, \ldots$ are infinitely many disjoint sets with nonempty interior, which are all contained in the compact set [0, 1].

[2] Actually, since there is no need to analyze probability distributions for this special case. One can prove corollary 3.1 of Casey (1996) more directly by considering the equivalence relation defined at the beginning of section 3, and by deriving a lower bound for the volume of the set of states that correspond to an equivalence class.

of a single network. However, there exists in general a small but nonzero probability that this average deviates strongly from its expected value. In addition, if one assumes that the computational unit that produces the output of the computation is also noisy, one cannot expect the reliability of the output of the computation to be larger than the reliability of this last computational unit. Consequently, there exist many methods for reducing the error probability of the output to a small value, but these methods cannot achieve error probability 0 at the output. In addition, if one wants to investigate computations with common noise distributions such as gaussian noise, which may in principle move a state to any other state, it is necessary to move to a computational model with less than perfect reliability of the output bit, since otherwise the model would not be able to carry out any nontrivial computations. Therefore, we focus our attention in this article on the general case where the reliability of the network output is just required to be $\geq 1/2 + \rho$ for some $\rho \in (0, \frac{1}{2}]$.

Unfortunately an investigation of computations with less than perfect reliability requires a more complex mathematical analysis. In a computational model with perfect reliability of the output, it cannot happen that an intermediate state q occurs at some step t in both a computation for an input x that leads to output 0 and at step t in a computation for the same input x that leads to output 1. Hence an analysis of perfectly reliable computations can focus on partitions of intermediate states q according to the computations and the computation steps where they may occur. In contrast, in a computational model with less than perfect reliability of the output bit, the same internal state q may occur at an intermediate step in computation paths that yield different output bits. Hence for such a model, one has to analyze probability distributions over intermediate states q.

Consider, for example, the special case of a sigmoidal neural net (with thresholding at the output), where for each input the output of an internal noisy sigmoidal gate is distributed according to some gaussian distribution (perhaps restricted to the range of all possible output values this sigmoidal gate can actually produce). In this case, an intermediate state q of the computational system is a vector of values produced by these gaussian distributions for different sigmoidal gates. Obviously each such intermediate state q can occur at any fixed step t in any computation (in particular in computations with different network output for the same network input). Hence perfect reliability of the network output is unattainable in this case. For an investigation of the actual computational power of a sigmoidal neural net with gaussian noise, one has to drop the requirement of perfect reliability of the output and instead analyze how probable it is that a particular network output is given and that a certain intermediate state is assumed. Hence, one has to analyze for each network input and each step t the different probability distributions over intermediate states q that are induced by computations of the noisy analog computational system. In fact, one may view the set of these probability distributions over intermediate states q

as a generalized set of states of a noisy analog computational system. In general the mathematical structure of this generalized set of states is substantially more complex than that of the original set of intermediate states q. In section 2, we define a rigorous mathematical model for this type of noisy analog computation and introduce some basic methods for analyzing this generalized set of states.

The preceding remarks may illustrate that if one drops the assumption of perfect reliability, then a more complex variety of computations becomes possible, and the computational power of a system may potentially increase. In fact, in theoretical computer science, a substantial number of constructions rely on the premise that the computational power of a digital computational system does in fact increase if it gets access to random bits and less than perfect reliability of the output bit is tolerated. This is relevant for the discussions of this article, since internal noise of a noisy computational system may also be viewed as something positive: as a free source of random numbers, which may actually be helpful for certain computations. In section 3 we prove an upper bound for the computational power of noisy analog computational systems that limits the potential impact of such effects in analog computation.

We show that under mild constraints on the noise characteristics, noisy analog systems with bounded finite-dimensional state-spaces have at most the computational power of finite automata. This upper bound is quite general, and it also covers practically relevant special cases such as systems with dependencies among different sources of stochasticity, as well as noisy computations in hybrid analog-digital computational models, such as a neural net combined with a binary register, or a network of noisy spiking neurons where a neuron may temporarily assume the discrete state not firing.

One goal of our investigation of the effects of analog noise is to find out which features of the noise process have the most detrimental effect on the computational power of an analog computational system. This turns out to be a nontrivial question. For example, one might think that analog noise that is likely to move an internal state over a large distance is more harmful than another type of analog noise that keeps an internal state within its neighborhood. However, this intuition is deceptive. Consider the extreme case of analog noise in a sigmoidal neural net that moves a gate output $x \in [-1, 1]$ to a value in some ε-neighborhood of $-x$, and compare it with noise that moves x to an arbitrary value in the 10ε-neighborhood of x. The first type of noise moves some values x over large distances but is likely to be less harmful for noise-robust computing than the second type, as the large jump from x to $-x$ represents just a recoding of the output value.

As a first step toward characterizing those aspects and parameters of analog noise that have a strong impact on the computational power of a noisy analog system, the proof of theorem 1 provides an explicit bound on the number of states of any finite automaton that can be implemented by an analog computational system with a given type of analog noise. It is quite surprising to see on which specific parameters of the analog noise the bound depends (c.f. the remark at the end of section 3).

In section 4 we prove a partial converse to the upper bound result in section 3 by showing that if one only considers bounded noise processes (where the analog noise can move an internal state at most over a distance η, for a sufficiently small value of η), then any finite automaton can be simulated with perfect ($\rho = 1/2$) reliability by a recurrent analog neural net of the type discussed in Anderson, Silverstein, Ritz, and Jones (1988) and Siegelmann and Sontag (1991). Other embeddings of finite automata in recurrent sigmoidal networks include Frasconi, Gori, Maggini, and Soda (1996) and Omlin and Giles (1996), which discuss, respectively, implementations of automata in noise-free radial basis function networks and in second-order networks with synaptic noise.

In section 5 we establish a new type of upper bound for the VC-dimension of computational models with analog noise. We show that in the presence of arbitrarily small amounts of analog noise, there exists an upper bound for the VC-dimension of, for example, neural nets that is independent of the total number of units in the case of a feedforward architecture, and independent of the computation time in the case of a recurrent neural net. This contrasts with the anomaly that in the noise-free setting, the classes of finite recurrent analog neural nets (Siegelmann & Sontag, 1991) and finite recurrent networks of spiking neurons (Maass, 1996) have infinite VC-dimension, and are thus strongly unlearnable from the point of view of learning theory. Again, the proofs of the theorem 3, and its corollaries 3 and 4, provide explicit (although very large) upper bounds for the VC-dimension of noisy analog neural nets with batch input, which depend on specific parameters of the analog noise.

2 Preliminaries: Computational Systems and Noise Processes

We shall define our computational model first in the noise-free setting and then consider the effect of noise on computations separately.

An analog discrete-time computational system (briefly: computational system) M is defined in a general way as a five-tuple $\langle \Omega, p^0, F, \Sigma, s \rangle$, where Ω, the set of states, is a bounded subset of $\mathbf{R}^d$, $p^0 \in \Omega$ is a distinguished initial state, $F \subseteq \Omega$ is the set of accepting states, Σ is the input domain, and $s : \Omega \times \Sigma \to \Omega$ is the transition function. To avoid unnecessary pathologies, we impose the conditions that Ω and F are Borel subsets of $\mathbf{R}^d$, and for each $a \in \Sigma$, $s(p, a)$ is a measurable function of p. We also assume that Σ contains a distinguished null value $\sqcup$, which may be used to pad the actual input to arbitrary length. The nonnull input domain is denoted by $\Sigma_0 = \Sigma - \{\sqcup\}$.

The intended noise-free dynamics of such a system M is as follows. The system starts its computation in state p^0, and on each single computation step on input, element $a \in \Sigma_0$ moves from its current state p to its next state $s(p, a)$. After the actual input sequence has been exhausted, M may still continue to make pure computation steps, which lead it from a state p to the state $s(p, \sqcup)$. The system accepts its input if it enters a state in the class F at some point after the input has finished. (We give a more precise definition of the dynamics, including the effects of noise, later.)

For instance, the recurrent analog neural net model of Siegelmann and Sontag (1991) (also known as the "brain state in a box" model of Anderson et al., 1988) is obtained from this general framework as follows. For a network $\mathcal{N}$ with d neurons and activation values between -1 and 1, the state-space is $\Omega = [-1, 1]^d$. The input domain may be chosen as either $\Sigma = \mathbf{R}$ or $\Sigma = \{-1, 0, 1\}$ (for online input) or $\Sigma = \mathbf{R}^n$ (for batch input). In each case the value zero (or the zero vector) serves conveniently as the null value $\sqcup$. For simplicity, we treat here formally only the cases where $\Sigma \subseteq \mathbf{R}$; the extensions to the case $\Sigma = \mathbf{R}^n$ are straightforward. The transition function $s : \Omega \times \Sigma \rightarrow \Omega$ is in this model given in terms of a $d \times d$ weight matrix $W = (w_{ij})$, a d-component bias vector $h = (h_i)$, a d-component input weight vector $c = (c_i)$, and a neuron activation function $\sigma : \mathbf{R} \rightarrow [-1, 1]$. For any $p \in \Omega$ and $a \in \Sigma$, we define $s(p, a) = p^+$, where for each $i = 1, \ldots, d$,

$$p_i^+ = \sigma\left(\sum_{j=1}^{d} w_{ij} p_j + h_i + c_i a\right).$$

Both Anderson et al. (1988) and Siegelmann and Sontag (1991) use the saturated-linear sigmoid activation function

$$\sigma(u) = \begin{cases} -1, & \text{if } u < -1, \\ u, & \text{if } -1 \leq u \leq 1, \\ 1, & \text{if } u > 1, \end{cases}$$

but one may obviously also define the model with respect to other activation functions, notably the standard sigmoid $\sigma(u) = \tanh u$, or the discontinuous signum function

$$\operatorname{sgn}(u) = \begin{cases} -1, & \text{if } u < 0, \\ 1, & \text{if } u \geq 0, \end{cases}$$

the latter choice yielding the model of recurrent threshold logic networks. The initial state in each of these models may be chosen as $p^0 = (-1, \ldots, -1)$, and the set of accepting states is determined by the activity of some specific output unit, say unit 1, so that $F = \{p \in \Omega \mid p_1 > \theta\}$, for some threshold value $\theta > 0$.

In the sequel, we shall use σ to denote the componentwise extension of the chosen activation function to state vectors, so that for any $p \in \Omega$, $\sigma(p) := (\sigma(p_1), \ldots, \sigma(p_d))$. This convention lets us write the transition function s defined above compactly as $s(p, a) = \sigma(Wp + h + ac)$.

Feedforward analog neural nets may also be modeled in the same manner, except that in this case, one may wish to select as the state set $\Omega := ([-1, 1] \cup \{\mathit{dormant}\})^d$, where *dormant* is a distinguished value not in $[-1, 1]$. This special value is used to indicate the state of a unit whose inputs have not all yet been available at the beginning of a given computation step (e.g., for units on the lth layer of a net at computation steps $t < l$).

The completely different model of a network of m stochastic spiking neurons (see, e.g., Gerstner & van Hemmen, 1994, or Maass, 1997) is also a special case of our general framework. In this case one wants to set $\Omega_{sp} := (\bigcup_{j=1}^{l} [0, T)^j \cup \{\mathit{not\text{-}firing}\})^m$, where $T > 0$ is a sufficiently large constant so that it suffices to consider only the firing history of the network during a preceding time interval of length T in order to determine whether a neuron fires (e.g., $T = 30$ ms for a biological neural system). If one partitions the time axis into discrete time windows $[0, T), [T, 2T), \ldots$, then in the noise-free case, the firing events during each time window are completely determined by those in the preceding one. A component $p_i \in [0, T)^j$ of a state in this set Ω_{sp} indicates that the corresponding neuron i has fired exactly j times during the considered time interval, and it also specifies the j firing times of this neuron during this interval. Due to refractory effects, one can choose $l < \infty$ for biological neural systems, for example, $l = 15$ for T = 30 ms. With some straightforward formal operations, one can also write this state set Ω_{sp} as a bounded subset of $\mathbf{R}^d$ for $d := l \cdot m$.

Let us then consider the effect of noise on computations. Let $Z(p, B)$ be a function that for each state $p \in \Omega$ and Borel set $B \subseteq \Omega$ indicates the probability of noise corrupting state p into some state in B. The function Z is called the noise process affecting M, and it should satisfy the mild conditions of being a stochastic kernel (Feller, 1971, p. 205), that is, for each $p \in \Omega$, $Z(p, \cdot)$ should be a probability distribution, and for each Borel set B, $Z(\cdot, B)$ should be a measurable function.

We assume that there is some measure μ over Ω so that $Z(p, \cdot)$ is absolutely continuous with respect to μ for each $p \in \Omega$; that is, $\mu(B) = 0$ implies $Z(p, B) = 0$ for every measurable $B \subseteq \Omega$. By the Radon–Nikodym theorem (Feller, 1971, p. 140), Z then possesses a density kernel with respect to μ; that is, there exists a function $z(\cdot, \cdot)$ such that for any state $p \in \Omega$ and Borel set $B \subseteq \Omega$,

$$Z(p, B) = \int_{q \in B} z(p, q)\, d\mu.$$

We assume that this function $z(\cdot, \cdot)$ has values in $[0, \infty)$ and is measurable. (Actually, in view of our other conditions, this can be assumed without loss of generality.)

The dynamics of a computational system M affected by a noise process Z is now defined as follows. If the system starts in a state p, the distribution of states q obtained after a single computation step on input $a \in \Sigma$ is given by the density kernel $\pi_a(p,q) = z(s(p,a),q)$. (Note that as a composition of two measurable functions, π_a is again a measurable function.) The long-term dynamics of the system is given by a Markov process, where the distribution $\pi_{xa}(p,q)$ of states after $|xa|$ computation steps with input $xa \in \Sigma^*$ starting in state p is defined recursively by

$$\pi_{xa}(p,q) = \int_{r\in\Omega} \pi_x(p,r) \cdot \pi_a(r,q)\, d\mu.$$

One easily can verify by induction on $|u|$ that

$$\pi_{xu}(p,q) = \int_{r\in\Omega} \pi_x(p,r) \cdot \pi_u(r,q)\, d\mu$$

for all $x, u \in \Sigma^*$ of length ≥ 1 .

Let us denote by $\pi_x(q)$ the distribution $\pi_x(p^0, q)$—the distribution of states of M after it has processed string x, starting from the initial state p^0. Let $\rho > 0$ be the required reliability level. In the most basic version, the system M accepts (rejects) some input $x \in \Sigma_0^*$ if $\int_F \pi_x(q)\, d\mu \geq \frac{1}{2} + \rho$ (respectively $\leq \frac{1}{2} - \rho$). In less trivial cases, the system may also perform pure computation steps after it has read all of the input. Thus, we define more generally that the system M recognizes a set $L \subseteq \Sigma_0^*$ with reliability ρ if for any $x \in \Sigma_0^*$:

$$x \in L \Leftrightarrow \int_F \pi_{xu}(q)\, d\mu \geq \frac{1}{2} + \rho \;\text{ for some } u \in \{\sqcup\}^*$$
$$x \notin L \Leftrightarrow \int_F \pi_{xu}(q)\, d\mu \leq \frac{1}{2} - \rho \;\text{ for all } u \in \{\sqcup\}^*.$$

This also covers the case of batch input, where $|x| = 1$ and Σ_0 is typically quite large (e.g., $\Sigma_0 = \mathbf{R}^n$).

One gets a reasonably realistic model for noise in an analog neural net with state-space $\Omega = [-1,1]^d$ by defining the noise process Z so that it reflects a clipped gaussian distribution. Without more specific knowledge about the noise source, this appears to be the most appropriate model for analog noise in an analog neural net. One assumes in this model that for any computation step, the intended output $p_i \in [-1,1]$ of the ith unit of the net is replaced by a clipped gaussian distribution of values $q_i \in [-1,1]$, where values < -1 (> 1) are rounded to -1 (respectively, 1). If one assumes that this rounding occurs independently for each of the d units i in the network and, for simplicity, that all the underlying gaussians have the same variance, then one arrives in our general framework for a noisy computational system M at a noise process Z where $Z(p,\cdot)$ is defined for each $p \in \Omega = [-1,1]^d$ by a symmetric gaussian distribution with density $z(p,q) = \nu(\| q - p \|)$ around p, but with all values $q_i < -1$ ($q_i > 1$) of the occurring states $\langle q_1, \ldots, q_d\rangle$ rounded to -1 (respectively 1). (Here $\| v \|$ denotes the Euclidean norm of a vector v, and ν is the density function of some symmetric d-variate gaussian distribution.) Since such a rounding process will assign probability > 0 to the lower-dimensional bounding hyperrectangles of Ω, we cannot simply define μ as the Lebesgue measure over Ω in order to subsume this type of analog noise under our general noise model. Rather one has to decompose Ω into components $\Omega_1, \ldots, \Omega_k$ (representing the interior Ω_1 and lower-dimensional bounding hyperrectangles $\Omega_2, \ldots, \Omega_k$ of $\Omega = [-1, 1]^d$), and define μ as a sum of measures $\mu_1 + \cdots + \mu_k$, where μ_1 is the Lebesgue measure over Ω_1 and $\mu_2, \ldots, \mu_k$ are Lebesgue measures for the lower-dimensional spaces $\Omega_2, \ldots, \Omega_k$.

In the case of a network of spiking neurons, the noise model has to take into account that not only the firing time of a neuron is subject to some jitter (which can be modeled by a gaussian distribution), but also neurons may randomly fail to fire, or they may fire "spontaneously" (even when they would not fire in the corresponding deterministic model). All these effects can be modeled by a suitable noise process Z defined on the state-space Ω_{sp} discussed earlier, with a measure μ over Ω defined by a decomposition of Ω similarly as in the case of analog neural nets.

3 An Upper Bound for the Computational Power of Systems with Analog Noise

It has been shown for various concrete models of analog computation without noise, such as generalized shift maps (Moore, 1990), recurrent neural nets (Siegelmann & Sontag, 1991), and networks of spiking neurons (Maass, 1996), that they can simulate a universal Turing machine, and hence have immense computational power. It has long been conjectured that their computational power collapses to that of a finite automaton as soon as one assumes that they are subject to even small amounts of analog noise. We provide in this section a proof of this conjecture. Furthermore we make explicit on which parameters of the analog noise the required number of states of a simulating finite automaton depends.

Our proof requires a mild continuity assumption for the density functions $z(r, \cdot)$, which is satisfied in all concrete cases that we have considered. We do not require any global continuity property over Ω for the density functions $z(r, \cdot)$ because of the previously discussed concrete cases, where the state-space Ω is a disjoint union of subspaces $\Omega_1, \ldots, \Omega_k$ with different measures on each subspace. We only assume that for some arbitrary partition of Ω into Borel sets $\Omega_1, \ldots, \Omega_k$ the density functions $z(r, \cdot)$ are uniformly continuous over each Ω_j, with moduli of continuity that can be bounded independent of r. In other words, we require that $z(\cdot, \cdot)$ satisfies the following condition:

A function $\pi(\cdot, \cdot)$ from Ω^2 into $\mathbf{R}$ is called piecewise equicontinuous if for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every $r \in \Omega$, and for all $p, q \in \Omega_j$, $j = 1, \ldots, k$:

$$\| p - q \| \leq \delta \quad \text{implies} \quad \left|\pi(r, p) - \pi(r, q)\right| \leq \varepsilon. \tag{3.1}$$

Note that because the state-space Ω is bounded, any restriction $\pi(r, \cdot)$ of a piecewise equicontinuous function $\pi(\cdot, \cdot)$ to fixed $r \in \Omega$ has bounded range. If $z(\cdot, \cdot)$ satisfies condition (3.1), we call also the resulting noise process Z piecewise equicontinuous. Our preceding discussions suggest that all practically relevant noise processes Z have this property.

To formulate our result, we need a notion of regular sets of sequences over arbitrary domains Σ_0 , which we define as follows. Let $L \subseteq \Sigma_0^*$ be a set of sequences over an input domain Σ_0. Sequences $x, y \in \Sigma_0^*$ are equivalent with respect to L if one has $xw \in L \Leftrightarrow yw \in L$ for all $w \in \Sigma_0^*$. The set L is regular if this equivalence relation has only finitely many equivalence classes. By the Myhill–Nerode theorem (Hopcroft & Ullman, 1979, pp. 65–67), for finite alphabets Σ_0, this definition coincides with the usual definition of regular sets via finite automata. From the point of view of computational complexity theory, machine models that accept only regular sets belong to the most "primitive" class of models. In contrast to Turing machines and other universal computational models, the number of internal states of such machine models is fixed, independent of the length of the input string.

Theorem 1. *Let $L \subseteq \Sigma_0^*$ be a set of sequences over an arbitrary input domain Σ_0. Assume that some computational system M, affected by a piecewise equicontinuous noise process Z, recognizes L with reliability ρ, for some arbitrary $\rho > 0$. Then L is regular.*

Proof. Let $M = \langle \Omega, p^0, F, \Sigma, s \rangle$, where $\Sigma = \Sigma_0 \cup \{\sqcup\}$, be the system in question recognizing L. We shall show that there are only finitely many equivalence classes of sequences with respect to L.

We begin by observing that if for two sequences $x, y \in \Sigma_0^*$, the distributions $\pi_x(\cdot)$ and $\pi_y(\cdot)$ are sufficiently close, then x and y are equivalent. To see this, assume that $\int_{r\in\Omega} \left|\pi_x(r) - \pi_y(r)\right| \, d\mu \leq \rho$, and suppose for a contradiction that x and y are not equivalent. Then there exists some $w \in \Sigma_0^*$ with $xw \in L \Leftrightarrow yw \notin L$. Without loss of generality, assume that $xw \in L$. Thus, there exists some $u \in \{\sqcup\}^*$ with $\int_F \pi_{xwu}(q)\, d\mu \geq \frac{1}{2} + \rho$ and $\int_F \pi_{ywu}(q)\, d\mu \leq \frac{1}{2} - \rho$. This yields the contradiction

$$\begin{aligned}
2\rho &\leq \left| \int_{q\in F} \pi_{xwu}(q)\, d\mu - \int_{q\in F} \pi_{ywu}(q)\, d\mu \right| \\
&= \left| \int_{q\in F} \int_{r\in\Omega} \pi_x(r) \cdot \pi_{wu}(r, q)\, d\mu\, d\mu - \int_{q\in F} \int_{r\in\Omega} \pi_y(r) \cdot \pi_{wu}(r, q)\, d\mu\, d\mu \right| \\
&\leq \int_{q\in F} \int_{r\in\Omega} \left|\pi_x(r) - \pi_y(r)\right| \cdot \pi_{wu}(r, q)\, d\mu\, d\mu \\
&= \int_{r\in\Omega} \left|\pi_x(r) - \pi_y(r)\right| \cdot \left(\int_{q\in F} \pi_{wu}(r, q)\, d\mu \right) d\mu
\end{aligned}$$

$$\leq \rho.$$

Thus we have shown that $\int_{r\in\Omega} \left|\pi_x(r) - \pi_y(r)\right| \, d\mu \leq \rho$ implies that $x, y \in \Sigma_0^*$ are equivalent.

Next we observe that all the density functions $\pi_x(\cdot)$ for $x \in \Sigma^*$ are piecewise uniformly continuous, with the same bounds on their moduli of continuity as the noise density functions $z(r, \cdot)$ have. This is verified by induction on $|x|$. Given $\varepsilon > 0$, let $\delta > 0$ be such that the density function $z(\cdot, \cdot)$ satisfies condition (3.1) for all $r \in \Omega$ and $j = 1, \ldots, k$. We then have for any $x \in \Sigma^+$, $a \in \Sigma$, and all $p, q \in \Omega_j$ such that $\| p - q \| \leq \delta$:

$$\begin{aligned}\left|\pi_{xa}(p) - \pi_{xa}(q)\right| &= \int_{r\in\Omega} \pi_x(r) \cdot \left|\pi_a(r, p) - \pi_a(r, q)\right| \, d\mu \\ &= \int_{r\in\Omega} \pi_x(r) \cdot \left|z(s(r, a), p) - z(s(r, a), q)\right| \, d\mu \\ &\leq \varepsilon \cdot \int_{r\in\Omega} \pi_x(r) \, d\mu \\ &= \varepsilon.\end{aligned}$$

The preceding observation now implies that the space of all functions $\pi_x(\cdot)$ for $x \in \Sigma_0^*$ can be partitioned into finitely many classes C so that any two functions $\pi_x(\cdot)$, $\pi_y(\cdot)$ in the same class C satisfy $\int_{r\in\Omega} \left|\pi_x(r) - \pi_y(r)\right| \, d\mu \leq \rho$, and hence correspond to sequences that are equivalent with respect to L. Such a partition can for example be achieved in the following way. Using the piecewise uniform continuity of the $\pi_x(\cdot)$, choose from within each component Ω_j of Ω a finite set (or "grid") G_j that is so dense that for each $r \in \Omega_j$, if $t_r \in G_j$ is the grid point closest to r, then $|\pi_x(r) - \pi_x(t_r)| \leq \rho/4\mu(\Omega)$. (To see that such a finite G_j always exists, note that given the value $\delta > 0$ corresponding to $\varepsilon = \rho/4\mu(\Omega)$ in condition 3.1, one can by the Bolzano-Weierstrass theorem choose only a finite number of points t from within the bounded set Ω_j so that any two distinct chosen points t, t' are more than a distance δ apart.) Take $G = \bigcup_{j=1}^k G_j$. Now partition the (bounded!) range of all functions $\pi_x(\cdot)$ into finitely many intervals I of length $\rho/2\mu(\Omega)$, and place two functions $\pi_x(\cdot)$, $\pi_y(\cdot)$ in the same class C if for every grid point $t \in G$ the values of $\pi_x(t)$ and $\pi_y(t)$ fall into the same interval I. Then for any two functions $\pi_x(\cdot)$, $\pi_y(\cdot)$ in the same class C it is the case that for any $r \in \Omega_j \subseteq \Omega$, $j = 1, \ldots, k$,

$$\begin{aligned}|\pi_x(r) - \pi_y(r)| &\leq |\pi_x(r) - \pi_x(t_r)| + |\pi_x(t_r) - \pi_y(t_r)| + |\pi_y(t_r) - \pi_y(r)| \\ &\leq \rho/\mu(\Omega),\end{aligned}$$

and thus $\int_{r\in\Omega} \left|\pi_x(r) - \pi_y(r)\right| \, d\mu \;\leq\; (\rho/\mu(\Omega)) \cdot \int_{r\in\Omega} d\mu = \rho.$

Remark. In stark contrast to the results of Siegelmann and Sontag (1991) and Maass (1996) for the noise-free case, the preceding theorem implies that both recurrent analog neural nets and recurrent networks of spiking neurons with online input from Σ_0^* can only recognize regular languages in the presence of any reasonable type of analog noise, even if their computation time is unlimited and if they employ arbitrary real-valued parameters.

Remark. The proof of theorem 1 relies on an analysis of the space of probability density functions over the state set Ω. An upper bound on the number of states of a deterministic finite automaton that simulates M can be given in terms of the number k of components Ω_j of the state set Ω, the dimension and diameter of Ω, a bound on the values of the noise density function z, and the value of δ corresponding to $\varepsilon = \rho/4\mu(\Omega)$ in condition 3.1.

4 Noisy Analog Neural Nets Recognize Regular Languages

Let us say that a noise process Z defined on a set $\Omega \subseteq \mathbf{R}^d$ is bounded by η if it can move a state p only to other states q that have a distance $\leq \eta$ from p in the L_1-norm over $\mathbf{R}^d$, that is, if its density kernel z has the property that for any $p = \langle p_1, \ldots, p_d \rangle$ and $q = \langle q_1, \ldots, q_d \rangle \in \Omega$, $z(p, q) > 0$ implies that $|q_i - p_i| \leq \eta$ for all $i = 1, \ldots, d$. As a partial converse to the upper-bound result of the previous section, we now prove that regular languages over the alphabet $\{-1, 1\}$ can be recognized with perfect reliability ($\rho = \frac{1}{2}$) by recurrent analog neural nets, as long as the noise process affecting the computation is bounded by a certain constant $\eta > 0$.

The basic idea of our proof is first to construct a threshold logic network $\mathcal{T}$ recognizing the regular language under consideration, and then simulate $\mathcal{T}$ with a noise-tolerant analog neural net. However, in order to obtain the tolerance versus delay trade-off results in a uniform manner, we derive them as corollaries from a general result on simulating threshold logic networks by noisy recurrent analog neural nets.

Consider a d-unit threshold logic network $\mathcal{T}$ (cf. section 2) with transition function $s(p, a) = \operatorname{sgn}(Wp + h + ac)$, where $W \in \mathbf{R}^{d\times d}$ is the weight matrix of $\mathcal{T}$, $h \in \mathbf{R}^d$ is the bias vector, and $c \in \mathbf{R}^d$ is the input weight vector. Let us say that $\mathcal{T}$ has separation θ, if at each unit, the argument to the signum function is always at least θ away from zero; that is, if $|w_i^T p + h_i + c_i a| \geq \theta$ always holds, for every $i = 1, \ldots, d$, $p \in \{-1, 1\}^d$, and $a \in \{-1, 0, 1\}$. Any threshold logic network operating on the input alphabet $\{-1, 0, 1\}$ may be modified to have some nonzero separation value by adjusting the bias values appropriately. An important special case are networks with integer weights, which may be adjusted to have separation 1. (On input values $a \in \{-1, 1\}$ this is straightforward; dealing with the value $a = 0$ may in some cases require modifying the network structure.)

Theorem 2. *Let a language $L \subseteq \{-1, 1\}^*$ be recognized by some d-unit threshold logic network $\mathcal{T}$ with separation $\theta > 0$, and let $w_{\max}$ be the maximum total input* *weight to any unit of $\mathcal{T}$ ($w_{\max} = \max_i \sum_j |w_{ij}|$). Let η be a constant satisfying $\eta < \theta/w_{\max}$. Then L can also be recognized by a d-unit recurrent analog neural net $\mathcal{N}$ that has perfect reliability ($\rho = \frac{1}{2}$) when affected by any noise process Z bounded by η. The activation function of $\mathcal{N}$ may be any function σ satisfying $\sigma(u) \to -1$ for $u \to -\infty$ and $\sigma(u) \to 1$ for $u \to \infty$.*

Proof. The idea of the proof is simply to simulate the threshold logic network $\mathcal{T}$ with an analog neural network $\mathcal{N}$ by forcing the analog units always to operate close to saturation (in states u such that $\sigma(u)$ is within δ of ± 1, for some small constant δ), so that they in effect function as threshold logic units. This is achieved by multiplying the weights in $\mathcal{N}$ by a sufficiently large constant m.

Thus, let a language L be recognized by a d-unit threshold logic network $\mathcal{T}$ with transition function $p^+ = s(p, a) = \operatorname{sgn}(Wp + h + ac)$, and separation θ.

Let δ and u_δ be constants such that the noise bound is $\eta < \theta/w_{\max} - \delta$, and for all $u \geq u_\delta$, $|1 - \sigma(u)| \leq \delta$, and for all $u \leq -u_\delta$, $|(-1) - \sigma(u)| \leq \delta$.

Now consider the analog network $\mathcal{N}$ obtained from $\mathcal{T}$ by multiplying all the weights and thresholds by a constant,

$$m \geq \frac{u_\delta}{\theta - w_{\max}(\eta + \delta)},$$

and replacing the signum nonlinearities by the sigmoids. We claim that $\mathcal{N}$ reproduces the behavior of $\mathcal{T}$ exactly, in the sense that the state of $\mathcal{N}$ at each time step, before noise is applied, is within δ of the corresponding state of $\mathcal{T}$.

Assume that the claim is true at some given time, when the state of $\mathcal{T}$ is some $p \in \{-1, 1\}^d$, and that of $\mathcal{N}$ correspondingly $\tilde{p} = p + r$, for some $r \in [-\delta, \delta]^d$. Consider then the update of $\mathcal{N}$ first with a noise vector $e = \tilde{q} - \tilde{p}$, where $\tilde{q}$ is generated according to some componentwise η-bounded noise density $z(\tilde{p}, \tilde{q})$, and then with the network transition function

$$\begin{aligned} \tilde{p}^+ &= \sigma(mW\tilde{q} + mh + mac) \\ &= \sigma(mW(p + r + e) + mh + mac) \\ &= \sigma(m(Wp + h + ac) + mW(r + e)). \end{aligned}$$

Considering the argument vector to the sigmoid componentwise, we obtain for each $i = 1, \ldots, d$ the bound:

$$|m(w_i^T p + h_i + c_i a) + m w_i^T (r + e)| \geq m\theta - m w_{\max}(\delta + \eta) \geq u_\delta.$$

By our choice of the value u_δ, we are thus again ensured that the components of the new state vector $\tilde{p}^+$ of $\mathcal{N}$ are within δ of the corresponding components of the state vector p^+ of $\mathcal{T}$. The claim follows by induction.

One technicality concerning the choice of final states in the network $\mathcal{N}$ still needs to be pointed out. Even though in the network $\mathcal{T}$ the final states may be defined as, say, $F_{\mathcal{T}} = \{p \in \{-1, 1\}^d \mid p_1 = 1\}$, noise in the network $\mathcal{N}$ also affects the state of the output unit, and so the final states there should be defined as $F_{\mathcal{N}} = \{p \in [-1, 1]^d \mid p_1 \geq 1 - \eta\}$, if the noise is bounded by η.

Corollary 1. *For every regular language $L \subseteq \{-1, 1\}^*$ there is a constant $\eta > 0$ such that L can be recognized with perfect reliability ($\rho = \frac{1}{2}$) by a recurrent analog neural net in spite of any noise process Z bounded by η.*

Proof. Let L be recognized by some finite automaton with m states. As presented in Minsky (1972, pp. 55–57), one can easily construct from this automaton a threshold logic network $\mathcal{T}$ with $2m + 1$ units that recognizes L. In Minsky's construction, there is one threshold logic unit for each *(state, input symbol)* pair of the simulated automaton, plus one unit that tests for the acceptance condition. (Actually, our model mandates testing also for input termination, which requires adding a few extra units.) A unit is activated (goes to state 1) when it receives an excitatory signal from some preceding *(state, symbol)* unit and its input line. All the nonzero weights in $\mathcal{T}$ have absolute value 1, and the units have fan-in at most $2m+1$. Since this network satisfies the conditions of theorem 2 with $\theta = 1$, $w_{\max} = 2m + 1$, we may choose any value of $\eta < 1/(2m + 1)$.

The next corollary shows that we can increase the noise tolerance of a network by slowing the computation. Given an integer constant $\tau \geq 1$, let us say that a network $\mathcal{N}$ recognizes a language L with delay τ, if for every string $x = a_1, \ldots, a_k \in \{-1, 1\}^*$, $x \in L$ if and only if $\mathcal{N}$ accepts the string $a_1^\tau, \ldots, a_k^\tau$ (each input symbol a_i is repeated τ times before the next one is presented).

Corollary 2. *For every regular language $L \subseteq \{-1, 1\}^*$ there is a constant delay value τ such that for any $\eta < \frac{1}{2}$, L can be recognized with delay τ with perfect reliability ($\rho = \frac{1}{2}$) by a recurrent analog neural net that may be subject to any noise process Z bounded by η.*

Proof. Let again L be recognized by some finite automaton with m states. The threshold logic units used in the simulation of corollary 1 simply test for the simultaneous activity on any one of the lines coming from the preceding *(state, symbol)* units and the appropriate input line. Thus, each such unit can be replaced by a tree of fan-in 2 OR gates, and a concluding AND gate. Considering that the maximum fan-in of the original units is $2m + 1$, the AND-OR trees may be constructed to have height $\tau = \lceil \log_2 m \rceil + 2$. The resulting network then has integer weights, with $w_{\max} = 2$, and recognizes the language L with delay τ.

Remark. One can obtain different noise tolerance versus delay trade-offs using the recent, more advanced simulations of finite automata by threshold logic networks (Alon, Dewdney, & Ott, 1991; Horne & Hush, 1996; Indyk, 1995). For instance, Horne and Hush (1996) presents a simulation of m-state finite automata by threshold logic networks with $O(\sqrt{m \log m})$ units, connection weights ±1, and delay 4. Thus, one can in corollary 2 achieve a noise-tolerance bound of $\eta = O(1/\sqrt{m \log m})$ with delay $\tau = 4$.

Remark. The precise values of the η bounds obtained above are proportional to the width of the interval used to encode unit states in the analog neural net model. The results are here formulated using the interval $[-1, 1]$, and changes in this interval would have the proportional effects on the η values. For instance, if the interval $[0, 1]$ were used (as in Siegelmann & Sontag, 1991), the η bound in corollary 2 would decrease from $\frac{1}{2}$ to $\frac{1}{4}$.

5 A Novel Upper Bound for the VC-Dimension of Various Types of Neural Nets with Analog Noise

In this section we provide an example for the effect of analog noise on discrete time analog computations with batch input. We focus our attention on the most common types of analog neural nets and show that in the presence of arbitrarily small amounts of analog noise, there exists an upper bound for the VC-dimension of such neural nets that is independent of the total number of gates in the case of a feedforward architecture and independent of the computation time in the case of a recurrent neural net. It depends on only the structure of the first layer of the neural net (or alternatively of any other fixed layer). This novel type of upper bound depends apart from the analog noise on only those parameters of the net that are relevant for its first computation step, and it holds for arbitrary real-valued batch inputs and arbitrary real-valued "programmable parameters" (weights, etc.).

The resulting upper bounds for the required sample size of a noisy multilayer sigmoidal neural net extend a preceding result by Haussler (1992). He had shown in corollary 3 that even in the noise-free case, an upper bound for the VC-dimension can be given that depends on only the maximal absolute value of weights for gates on layers ≥ 2 and on their maximal fan-in. In the present result, all dependence on parameters that concern gates on layers ≥ 2 is removed.

It will become obvious from the proof of theorem 3 that our upper bound is actually of a quite general nature, and it can also be applied to various other models for discrete-time analog computation with analog noise that are not related to neural nets.

The VC-dimension (abbreviated VC-dim($\mathcal{F}$)) of an arbitrary class $\mathcal{F}$ of functions $f : \mathbf{R}^n \to \{0, 1\}$ is defined as follows. One says that $\mathcal{F}$ shatters a finite set $S \subseteq \mathbf{R}^n$ if for every subset $A \subseteq S$ there exists a function $f \in \mathcal{F}$ with

$f(x) = 1$ for $x \in A$ and $f(x) = 0$ for $x \in S - A$. The VC-dimension of $\mathcal{F}$ is defined as VC-dim$(\mathcal{F}) :=$ sup $\{|S| : S \subseteq \mathbf{R}^n$ is shattered by $\mathcal{F}\}$.

The VC-dimension of $\mathcal{F}$ may be viewed as a measure for the expressibility (or degrees of freedom) of $\mathcal{F}$. In particular, it provides for arbitrary finite sets $D \subseteq \mathbf{R}^n$ an upper bound of the form $|D|^{O(\text{VC-dim}(\mathcal{F}))}$ for the number of functions $D \to \{0, 1\}$ that can be written as a restriction of a function in $\mathcal{F}$ to this finite domain D. As a consequence, the VC-dimension of $\mathcal{F}$ is the key parameter for estimating the number of randomly chosen examples that are needed to "learn" arbitrary target functions $g : \mathbf{R}^n \to \{0, 1\}$ from randomly chosen examples $\langle x, g(x)\rangle$ for g by a learning algorithm that uses functions from $\mathcal{F}$ as hypotheses (see Haussler, 1992; Vapnik & Chervonenkis, 1971; Blumer, Ehrenfeucht, Haussler, & Warmuth, 1989; Maass, 1995). It should be noted that this does not only hold for the "classical" probably approximately correct (PAC) learning model where the target function g is required to belong to the class $\mathcal{F}$, but according to Haussler (1992), also in the general case of agnostic PAC learning where $g : \mathbf{R}^n \to \{0, 1\}$ can be any function. Of course, the latter case is much more relevant for the theory of learning with neural nets, where the class $\mathcal{F}$ of possible "hypotheses" is fixed by the architecture of the neural net on which we run a learning algorithm, whereas the examples $\langle x, g(x)\rangle$ may arise from some arbitrary real-world classification problem for which we train the neural net.

It is obvious from the results of Siegelmann and Sontag (1991) and Maass (1996) that there exist finite recurrent analog neural nets and finite recurrent networks of spiking neurons with batch input and parameters from $\mathbf{Q}$ that have infinite VC-dimension (consider networks that can simulate a universal Turing machine, with each input bit-string encoded into a rational number). From the point of view of learning theory, an infinite VC-dimension is commonly interpreted as information-theoretic evidence that there exists no "learning algorithm" for such networks (not even one with unlimited computation time). We will show in this section that this "anomaly" disappears as soon as one takes into account that the neural net is subject to analog noise, even if the amount of such noise is arbitrarily small.

For technical reasons, we also discuss the pseudo-dimension P-dim$(\mathcal{G})$ of a class $\mathcal{G}$ of real-valued functions $g : \mathbf{R}^n \to \mathbf{R}$. One can define P-dim$(\mathcal{G})$ as the VC-dimension of the following associated class,

$$\begin{aligned}\mathcal{F} :&= \{f\colon \mathbf{R}^{n+1} \to \{0, 1\}\colon \exists\, g \in \mathcal{G}(f(x, y) = 1 \\ &\quad \text{if } g(x) \geq y \text{ and } f(x, y) = 0 \text{ if } g(x) < y)\},\end{aligned}$$

of boolean-valued functions.

Consider now the computation of a system $M = \langle \Omega, p^0, F, \mathbf{R}^n, s\rangle$ on a batch input vector $x \in \mathbf{R}^n$, affected by some piecewise equicontinuous noise process Z whose density function z has values in some range $[0, B]$. The distribution of states of M after $k \geq 1$ computation steps is given by the density function $\pi_{xu}(p)$, where $|x| = 1$ and $u = \sqcup^{k-1}$. For $k > 1$, this density can be decomposed as $\int_{q\in\Omega}\pi_x(p^0,q)\cdot\pi_u(q,p)\,d\mu$, and for $k=1$ we have simply $\pi_{xu}(p)=\pi_x(p^0,q)$. This decomposition of the density function for the state-distribution of M will be essential for our subsequent results. We show in theorem 3 that there exists a finite upper bound for the VC-dimension of the class $\mathcal{F}$ of functions computable by a class $\mathcal{M}$ of such systems M (which receive arbitrary real-valued batch input) that does not depend on the complexity of the class $\mathcal{H}$ of functions $\pi_z(\cdot,\cdot)$ that describe the second part of the computations of these systems M after their first computation step.

Let $\mathcal{M}$ be a class of such systems, affected by the same piecewise equicontinuous noise process Z. For example, $\mathcal{M}$ can be the class of systems M that result from different weight assignments to some feedforward or recurrent analog neural net with some fixed architecture. Denote by $\mathcal{G}$ the class of all density kernels of the form $\pi(x,q):=\pi_x(p^0,q)$ for systems $M\in\mathcal{M}$, and by $\mathcal{H}$ the class of density kernels of the form $\omega(q,p):=\pi_u(q,p)$, for systems $M\in\mathcal{M}$ and sequences $u\in\{\sqcup\}^*$. (As a special case, we include also the constant function 1 in $\mathcal{H}$.) Then all the boolean functions computed with reliability ρ by the systems $M\in\mathcal{M}$ are included in the class $\mathcal{F}$ of functions $f:\mathbf{R}^n\to\{0,1\}$ that are composed of a function $\pi\in\mathcal{G}$ and a function $\omega\in\mathcal{H}$ so that for any $x\in\mathbf{R}^n$ the integral

$$\int_{p\in F}\int_{q\in\Omega}\pi(x,q)\cdot\omega(q,p)\,d\mu\,d\mu \text{ has a value } \geq \tfrac{1}{2}+\rho \text{ if } f(x)=1, \\ \text{and else a value } \leq \tfrac{1}{2}-\rho. \tag{5.1}$$

Actually, the class $\mathcal{F}$ contains somewhat more than just the functions computed by systems from $\mathcal{M}$, because the two component functions π and ω in equation 5.1 may come from two different systems in $\mathcal{M}$ (for example, from two different weight assignments to a recurrent analog neural net).

In theorem 3 we consider an even more general setup where one has two bounded state sets $\Omega\subseteq\mathbf{R}^d$ and $\Omega'\subseteq\mathbf{R}^{d'}$, measures μ over Ω and μ' over Ω', as well as a Borel set $F\subseteq\Omega'$ of accepting final states. (In applications Ω is typically the set of possible intermediates states after a fixed number l (e.g., $l=1$) of computation steps, and Ω' is the set of possible output states of a computation. One has $d\neq d'$ if, for example, the number d of units on the first hidden layer of a feedforward sigmoidal neural net differs from the number d' of output nodes of the net; see corollary 3.)

We assume in theorem 3 that $\mathcal{G}$ is an arbitrary class of piecewise equicontinuous density kernels $\pi:\mathbf{R}^n\times\Omega\to[0,B]$ with uniformly bounded moduli of continuity (as in condition 3.1), that $\mathcal{H}$ is an arbitrary class of density kernels $\omega:\Omega\times\Omega'\to\mathbf{R}^+$, that $\rho>0$ is an arbitrary given parameter, and that $\mathcal{F}$ is the class of functions $f:\mathbf{R}^n\to\{0,1\}$ for which there exist functions $\pi\in\mathcal{G}$ and $\omega\in\mathcal{H}$ so that for any $x\in\mathbf{R}$ the integral $\int_{p\in F}\int_{q\in\Omega}\pi(x,q)\cdot\omega(q,p)\,d\mu\,d\mu'$ has a value $\geq\frac{1}{2}+\rho$ if $f(x)=1$, and otherwise a value $\leq\frac{1}{2}-\rho$.

Because of our assumption about the function class $\mathcal{G}$, one can (as in the proof of theorem 1) superimpose on the space Ω a finite grid G, such that for any $\pi, \tilde{\pi} \in \mathcal{G}$ and $x, \tilde{x} \in \mathbf{R}^n$: $|\pi(x, q) - \tilde{\pi}(\tilde{x}, q)| \leq \rho/5\mu(\Omega)$ for all $q \in G$ implies that $\int_{q\in\Omega} \left|\pi(x, q) - \tilde{\pi}(\tilde{x}, q)\right| \, d\mu < \rho/2$. The size $|G|$ of the grid (that is, the number of grid points) depends in general on the reliability parameter ρ, the common moduli of continuity of the functions in $\mathcal{G}$, and the volume and shape of the state-space Ω.

Theorem 3. *Let $\mathcal{G}$, $\mathcal{H}$, and $\mathcal{F}$ be function classes as specified above and assume in addition that the class $\mathcal{G}$ has finite pseudo-dimension Δ. Then one can give a finite upper bound for the VC-dimension of $\mathcal{F}$ in terms of ρ, B, $|G|$, Δ, and $\mu(\Omega)$. Obviously this bound does not depend on the complexity of the function class $\mathcal{H}$ (except via parameters related to the state set Ω) .*

Proof. Let $S \subseteq \mathbf{R}^n$ be some arbitrary finite set shattered by $\mathcal{F}$. For any subset $A \subseteq S$ we fix functions $\pi_A \in \mathcal{G}$ and $\omega_A \in \mathcal{H}$ so that for any $x \in S$ the integral $\int_{p\in F} \int_{q\in\Omega} \pi_A(x, q) \cdot \omega_A(q, p) \, d\mu \, d\mu'$ has a value $\geq \frac{1}{2} + \rho$ if $x \in A$, and else a value $\leq \frac{1}{2} - \rho$. We write $\mathcal{G}_S^*$ for the class of all functions $\pi_A \in \mathcal{G}$ for $A \subseteq S$, and $\mathcal{G}_S$ for the class of restrictions of these functions to the finite domain $S \times G$.

We also consider for $\gamma := \rho/10\mu(\Omega)$ and any class $\mathcal{A}$ of functions with range $\mathbf{R}^+$ the class $\mathcal{A}^\gamma$ of all "γ-discretizations" g^γ of functions $g \in \mathcal{A}$, where

$$g^\gamma(z) := \left\lfloor \frac{g(z)}{\gamma} \right\rfloor \text{ for any } z \text{ in the domain of } g.$$

In particular for the class $\mathcal{G}_S$ the functions $\pi_A^\gamma \in \mathcal{G}_S^\gamma$ map $S \times G$ into $\{0, \ldots, b-1\}$ for $b := \lfloor B/\gamma \rfloor + 1$. Note that by our assumptions on $\mathcal{G}$, for any $\pi, \tilde{\pi} \in \mathcal{G}$ and any $x, \tilde{x} \in S$ the condition $\forall q \in G\ (|\pi^\gamma(x, q) - \tilde{\pi}^\gamma(\tilde{x}, q)| \leq 1)$ implies that $\int_\Omega |\pi(x, q) - \tilde{\pi}(\tilde{x}, q)| \, d\mu < \rho/2$.

One can get an upper bound for the complexity of $\mathcal{G}_S^*$ by applying to $\mathcal{G}_S^\gamma$ a generalization of Sauer's lemma due to Alon, Cesa-Bianci, Ben-David, and Haussler (1993). Given integers m, b, and Δ, define $\beta(m, b, \Delta) := \log_2 \sum_{i=1}^{\Delta} \binom{m}{i} b^i$. Lemma 15 of Alon et al. (1993) states that if $\mathcal{A}^\gamma$ is any class of functions obtained as the discretizations of the functions in a class $\mathcal{A}$ of pseudo-dimension Δ, such that the functions in $\mathcal{A}^\gamma$ have a domain D of size m and range $\{0, \ldots, b-1\}$, then $\mathcal{A}^\gamma$ must contain an "L_∞ 2-cover" $\mathcal{B}^\gamma \subseteq \mathcal{A}^\gamma$ of size at most $|\mathcal{B}^\gamma| \leq 2 \cdot (mb^2)^{\beta(m,b,\Delta)}$. That is, for every $f \in \mathcal{A}^\gamma$ there is some $\tilde{f} \in \mathcal{B}^\gamma$ such that $|f(z) - \tilde{f}(z)| < 2$ (and hence ≤ 1) for every $z \in D$. (The result holds for general values of γ, Δ, m, and b.)

Applied to our context (with $\mathcal{A} := \mathcal{G}_S$), this result implies that there exists a set $\mathcal{G}^* \subseteq \mathcal{G}_S^*$ whose cardinality can be bounded in terms of the pseudo-dimension Δ of $\mathcal{G}$ as

$$|\mathcal{G}^*| \leq 2 \cdot (|S| \cdot |G| \cdot b^2)^{\beta(|S|\cdot|G|, b, \Delta)}, \tag{5.2}$$

such that for every $\pi \in \mathcal{G}_S^*$ there exists some $\tilde{\pi} \in \mathcal{G}^*$ with $|\pi^\gamma(x,q) - \tilde{\pi}^\gamma(x,q)| \leq 1$ for all $x \in S$ and all $q \in G$.

With the help of the 2-cover of $\mathcal{G}_S^\gamma$ induced by $\mathcal{G}^*$, we can now show that the cardinality $|S|$ of the shattered set S can be bounded through the inequality

$$2^{|S|} \leq 2 \cdot (|S| \cdot |G| \cdot b^2)^{\beta(|S|\cdot|G|,b,\Delta)} \cdot 2^{b^{|G|}}. \tag{5.3}$$

It is obvious that this inequality yields an upper bound for $|S|$ that does not depend on the complexity of the function class $\mathcal{H}$ (except for parameters related to Ω).

Let us consider for each $\omega \in \mathcal{H}$ the discrete map $\hat{\omega} : \{0, \ldots, b-1\}^G \rightarrow \{0,1\}$ which is induced by ω through the following definition: $\hat{\omega}(\hat{\pi})$ has value 0 for $\hat{\pi} \in \{0, \ldots, b-1\}^G$ if there exist some $\pi \in \mathcal{G}$ and $x \in S$ with $|\hat{\pi}(q) - \pi^\gamma(x,q)| \leq 1$ for all $q \in G$ and $\int_{p\in F}\int_{q\in\Omega} \pi(x,q)\cdot\omega(q,p)\, d\mu\, d\mu' \leq \frac{1}{2} - \rho$. Else we set $\hat{\omega}(\hat{\pi}) = 1$.

Since we have the upper bound (see equation 5.2) on the size of the cover $\mathcal{G}^*$, and there exist at most $2^{b^{|G|}}$ different functions $\hat{\omega}$, it suffices for proving equation 5.3 to show that the following claim holds.

Claim. Let $A_1, A_2 \subseteq S$. If some function $\tilde{\pi} \in \mathcal{G}^*$ covers both π_{A_1} and π_{A_2}, in the sense that $|\tilde{\pi}^\gamma(x,q) - \pi_{A_1}^\gamma(x,q)| \leq 1$ and $|\tilde{\pi}^\gamma(x,q) - \pi_{A_2}^\gamma(x,q)| \leq 1$ for all $x \in S$ and all $q \in G$, and moreover $\hat{\omega}_{A_1} = \hat{\omega}_{A_2}$, then $A_1 = A_2$. In order to prove this claim, let us assume that $A_1 \neq A_2$, but both π_{A_1} and π_{A_2} are covered by the same function $\tilde{\pi} \in \mathcal{G}^*$. We shall show that $\hat{\omega}_{A_1} \neq \hat{\omega}_{A_2}$.

Fix some $x_0 \in S$ so that either $x_0 \in A_1 - A_2$ or $x_0 \in A_2 - A_1$. Without loss of generality, we may assume that $x_0 \in A_1 - A_2$. Let $\hat{\pi} : G \rightarrow \{0, \ldots, b-1\}$ be defined by $\hat{\pi}(q) = \tilde{\pi}^\gamma(x_0, q)$. Then we have $\hat{\omega}_{A_2}(\hat{\pi}) = 0$, since by assumption $|\tilde{\pi}^\gamma(x_0,q) - \pi_{A_2}^\gamma(x_0,q)| \leq 1$ for all $q \in G$ and

$$\int_{p\in F}\int_{q\in\Omega} \pi_{A_2}(x_0,q)\cdot\omega_{A_2}(q,p)\, d\mu\, d\mu' \leq \frac{1}{2} - \rho.$$

Assume for a contradiction that also $\hat{\omega}_{A_1}(\hat{\pi}) = 0$ for this function $\hat{\pi}$. This implies that there exist some $\pi \in \mathcal{G}$ and some $x_1 \in S$ with

$$\int_{p\in F}\int_{q\in\Omega} \pi(x_1,q)\cdot\omega_{A_1}(q,p)\, d\mu\, d\mu' \leq \frac{1}{2} - \rho \quad \text{and} \tag{5.4}$$
$$|\hat{\pi}(q) - \pi^\gamma(x_1,q)| \leq 1 \text{ for all } q \in G.$$

The latter implies by our choice of G and γ and the definition of $\hat{\pi}$ that

$$\int_{q\in\Omega} |\tilde{\pi}(x_0,q) - \pi(x_1,q)|\, d\mu < \rho/2. \tag{5.5}$$

On the other hand, the assumptions on $\tilde{\pi} \in \mathcal{G}^*$ imply that $|\tilde{\pi}^{\gamma}(x_0, q) - \pi_{A_1}^{\gamma}(x_0, q)| \leq 1$ for all $q \in G$, hence

$$\int_{q\in\Omega} |\tilde{\pi}(x_0, q) - \pi_{A_1}(x_0, q)|\, d\mu < \rho/2\,. \tag{5.6}$$

Furthermore since $x_0 \in A_1$, we have by choice of ω_{A_1} that

$$\int_{p\in F}\int_{q\in\Omega} \pi_{A_1}(x_0, q)\cdot\omega_{A_1}(q, p)\, d\mu\, d\mu' \geq \frac{1}{2} + \rho. \tag{5.7}$$

The inequalities (equations 5.5 and 5.6) imply that

$$\int_{q\in\Omega} |\pi(x_1, q) - \pi_{A_1}(x_0, q)|\, d\mu < \rho.$$

This inequality yields in combination with equations 5.4 and 5.7 the contradiction

$$\begin{aligned}
\rho &\leq \left| \int_{p\in F}\int_{q\in\Omega} \pi(x_1, q)\cdot\omega_{A_1}(q, p)\, d\mu\, d\mu' \right.\\
&\qquad \left. - \int_{p\in F}\int_{q\in\Omega} \pi_{A_1}(x_0, q)\cdot\omega_{A_1}(q, p)\, d\mu\, d\mu' \right|\\
&\leq \int_{p\in F}\int_{q\in\Omega} \left|\pi(x_1, q) - \pi_{A_1}(x_0, q)\right|\cdot\omega_{A_1}(q, p)\, d\mu\, d\mu'\\
&= \int_{q\in\Omega} \left|\pi(x_1, q) - \pi_{A_1}(x_0, q)\right|\cdot\left(\int_{p\in F}\omega_{A_1}(q, p)\, d\mu'\right) d\mu\\
&\leq \int_{q\in\Omega} \left|\pi(x_1, q) - \pi_{A_1}(x_0, q)\right|\, d\mu\\
&< \rho.
\end{aligned}$$

This contradiction implies that $\hat{\omega}_{A_1}(\hat{\pi}) = 1$, hence $\hat{\omega}_{A_1} \neq \hat{\omega}_{A_2}$. Thus we have verified the preceding claim, and the proof of theorem 3 is now complete.

Remark. It follows from Alon et al. (1993) that instead of a finite upper bound for the pseudo-dimension of $\mathcal{G}$, it suffices for theorem 3 to assume a finite upper bound for the γ-dimension P_γ-dim$(\mathcal{G})$ of $\mathcal{G}$ for $\gamma = \rho/20\mu(\Omega)$.

Corollary 3. *There exists a finite upper bound for the VC-dimension of layered feedforward sigmoidal neural nets and feedforward networks of spiking neurons with piecewise equicontinuous analog noise (for arbitrary real-valued inputs, boolean output computed with some arbitrary reliability $\rho > 0$, and arbitrary real-valued "programmable parameters") that does not depend on the size or structure of the network beyond its first hidden layer.*

Proof. We first consider for some arbitrary given parameters $n, d, d' \in \mathbf{N}$ the class $\mathcal{N}$ of all layered feedforward sigmoidal neural nets with n input nodes, d units on their first hidden layer, and d' output nodes. Thus, the nets in $\mathcal{N}$ may have arbitrary numbers of layers and gates and arbitrary real-valued weight assignments. We assume that the d gates on the first layer are affected by some piecewise equicontinuous noise process with density kernel $z : \Omega^2 \to \mathbf{R}^+$, where $\Omega := [-1, 1]^d$. Let $F : \mathbf{R}^n \times \mathbf{R}^m \to \Omega$ be the function whose value $F(x, w)$ is the vector of outputs of the d first hidden-layer units (without noise), for arbitrary network inputs $x \in \mathbf{R}^n$ and arbitrary assignments $w \in \mathbf{R}^m$ to the weights and biases of these units.

We take as the class $\mathcal{G}$ of functions π considered in the proof of Theorem 3 all functions of the form $\pi(x, q) = z(F(x, w), q)$ for arbitrary parameters $w \in \mathbf{R}^m$. The results presented in Karpinski and Macintyre (1997) imply that the pseudo-dimension of this class $\mathcal{G}$ of functions is bounded by a polynomial in m, for all common choices of activation functions of the sigmoidal units and all practically relevant density kernels z for the noise process (even involving the exponential function). In the case where the activation functions and density kernels are piecewise polynomial, one can apply the results of Goldberg and Jerrum (1995) to get a slightly better finite upper bound for the pseudo-dimension of $\mathcal{G}$.

We define for $\Omega = [-1, 1]^d$ and $\Omega' = [-1, 1]^{d'}$ the class $\mathcal{H}$ as the class of all density kernels $\omega : \Omega \times \Omega' \to \mathbf{R}^+$ that describe the computations of the remaining layers of networks in $\mathcal{N}$ with arbitrary noise processes (and arbitrary real-valued weights).

It follows from theorem 3 that the finite VC-dimension bound obtained for the class $\mathcal{F}$ of functions computed with reliability $\rho > 0$ by networks in the class $\mathcal{N}$ does not depend on the complexity of the function class $\mathcal{H}$, and hence not on the number of layers, the number of units beyond the first layer, or the noise process on later layers of these networks.

In the case of a network N of noisy spiking neurons, the programmable parameters consist of the “weights” of synapses, time delays for postsynaptic potentials, and parameters that determine other aspects of the functional form of response functions (i.e., postsynaptic potentials) and threshold functions. The pseudo-dimension of the class $\mathcal{G}$ that arises when one applies (as described in section 2) the framework considered here to the first layer of a network N of noisy spiking neurons can be bounded with the help of the same tools as for the case of sigmoidal neural nets.

Corollary 4. *There exists a finite upper bound for the VC-dimension of recurrent sigmoidal neural nets and networks of spiking neurons with analog noise (for arbitrary real valued inputs, boolean output computed with some arbitrary reliability $\rho > 0$, and arbitrary real valued “programmable parameters”) that does not depend on the computation time of the network, even if the computation time is allowed to vary for different inputs.*

Proof. One proceeds in the same manner as for the proof of corollary 3, except that $\mathcal{G}$ now consists of the class of all state distributions that arise from the first computation step of the total network, and $\mathcal{H}$ consists of all possible state transformations that can arise from the rest of the computations of the same network.

6 Conclusions

We have introduced a new framework for the analysis of analog noise in discrete-time analog computations that is better suited for real-world applications and more flexible than previous models. In contrast to preceding models, it also covers important concrete cases such as analog neural nets with a gaussian distribution of noise on analog gate outputs, noisy computations with less than perfect reliability, and computations in networks of noisy spiking neurons.

Furthermore, we have introduced adequate mathematical tools for analyzing the effect of analog noise in this new framework. These tools differ quite strongly from those that have been used previously for the investigation of noisy computations. We show that they provide new bounds for the computational power and VC-dimension of analog neural nets and networks of spiking neurons in the presence of analog noise.

Finally, our model for noisy analog computations can also be applied to completely different types of models for discrete-time analog computation than neural nets, such as arithmetical circuits (Turán and Vatan, 1994), the random access machine with analog inputs, the parallel random access machine with analog inputs, various computational discrete-time dynamical systems (Moore, 1990; Koiran, Cosnard, & Garzon, 1994; Asarin & Maler, 1994; Orponen & Matamala, 1996) and (with some minor adjustments) also the BSS model (Blum, Shub, & Smale, 1989; Koiran, 1993). Our framework provides for each of these models an adequate definition of noise-robust computation in the presence of analog noise, and our results provide upper bounds for their computational power and VC-dimension in terms of characteristics of their analog noise.

Acknowledgments

We thank Peter Auer for helpful conversations.

Part of this work was done while P. O. was at the University of Helsinki and during visits to the Technische Universität Graz and the University of Chile in Santiago.

This work was supported by the "Fonds zur Förderung der wissenschaflichen Forschung (FWF), Austrian Science Fund," project number P12153.

References

Alon, N., Dewdney, A. K., & Ott, T. J. (1991). Efficient simulation of finite automata by neural nets. *J. Assoc. Comput. Mach., 38*, 495–514.

Alon, N., Cesa-Bianchi, N., Ben-David, S., & Haussler, D. (1993). Scale-sensitive dimensions, uniform convergence, and learnability. In *Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science* (pp. 292–301). New York: IEEE Computer Science Press.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1988). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. In J. A. Anderson & E. Rosenfeld (Eds.), *Neurocomputing: Foundations of research.* Cambridge, MA: MIT Press.

Asarin, E., & Maler, O. (1994). On some relations between dynamical systems and transition systems. In *Proceedings of the 21st International Colloquium on Automata, Languages, and Programming* (pp. 59–72). Lecture Notes in Computer Science 820. Berlin: Springer-Verlag.

Blum, L., Shub, M., & Smale, S. (1989). On a theory of computation over the real numbers: NP-completeness, recursive functions and universal machines. *Bulletin of the Amer. Math. Soc., 21*, 1–46.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *J. Assoc. Comput. Mach., 36*, 929–965.

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

Cowan, J. D. (1966). Synthesis of reliable automata from unreliable components. In E. R. Caianiello (Ed.), *Automata theory* (pp. 131–145). New York: Academic Press.

Feller, W. (1971). *An introduction to probability theory and its applications* (2nd ed.). New York: Wiley.

Frasconi, P., Gori, M., Maggini, M., & Soda, G. (1996). Representation of finite state automata in recurrent radial basis function networks. *Machine Learning, 23*, 5–32.

Gál, A. (1991). Lower bounds for the complexity of reliable boolean circuits with noisy gates. In *Proceedings of the 32th Annual IEEE Symposium on Foundations of Computer Science* (pp. 594–601). New York: IEEE Computer Science Press.

Gerstner, W., & van Hemmen, J. L. (1994). How to describe neuronal activity: Spikes, rates or assemblies? In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems, 6* (pp. 463–470). San Mateo, CA: Morgan Kaufmann.

Goldberg, P. W., & Jerrum, M. R. (1995). Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning, 18*, 131–148.

Haussler, D. (1992). Decision theoretic generalizations of the PAC-model for neural nets and other learning applications. *Information and Computation, 100*, 78–150.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages, and computation*. Reading, MA: Addison-Wesley.

Horne, B. G., & Hush, D. R. (1996). Bounds on the complexity of recurrent neural network implementations of finite state machines. *Neural Networks, 9*, 243–252.

Indyk, P. (1995). Optimal simulation of automata by neural nets. In *Proceedings of the 12th Annual Symposium on Theoretical Aspects of Computer Science* (pp. 337–347). Lecture Notes in Computer Science 900. Berlin: Springer-Verlag.

Karpinski, M., & Macintyre, A. (1997). Polynomial bounds for VC-dimension of sigmoidal and general Pfaffian neural networks. *J. Computer and System Sciences, 54*, 169–179.

Kifer, Y. (1988). *Random perturbations of dynamical systems*. Boston: Birkhäuser.

Koiran, P. (1993). A weak version of the Blum, Shub and Smale model. In *Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science* (pp. 486–495). New York: IEEE Computer Science Press.

Koiran, P., Cosnard, M., & Garzon, M. (1994). Computability with low-dimensional dynamical systems. *Theoret. Comput. Sci., 132*, 113–128.

Maass, W. (1995). Vapnik-Chervonenkis dimension of neural nets. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (pp. 1000–1003). Cambridge, MA: MIT Press.

Maass, W. (1996). Lower bounds for the computational power of networks of spiking neurons. *Neural Computation, 8*, 1–40.

Maass, W. (1997). Fast sigmoidal networks via spiking neurons. *Neural Computation, 9*, 279–304.

Minsky, M. L. (1972). *Computation: Finite and infinite machines.* Englewood Cliffs, NJ: Prentice Hall.

Moore, C. (1990). Unpredictability and undecidability in physical systems. *Phys. Review Letters, 64*, 2354–2357.

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

Orponen, P., & Matamala, M. (1996). Universal computation by finite two-dimensional coupled map lattices. In *Proceedings of the Workshop on Physics and Computation, PhysComp'96* (pp. 243–247). Boston: New England Complex Systems Institute.

Phatak, D. S., & Koren, I. (1995). Complete and partial fault tolerance of feedforward neural nets. *IEEE Transactions on Neural Networks, 6*, 446–456.

Pippenger, N. (1989). Invariance of complexity measures for networks with unreliable gates. *J. Assoc. Comput. Mach., 36*, 531–539.

Rabin, M. (1963). Probabilistic automata. *Information and Control, 6*, 230–245.

Siegelmann, H. T. (1994). On the computational power of probabilistic and faulty networks. In *Proceedings of the 21st International Colloquium on Automata, Languages, and Programming* (pp. 23–34). Lecture Notes in Computer Science 820. Berlin: Springer-Verlag.

Siegelmann, H. T., & Sontag, E. D. (1991). Turing computability with neural nets. *Appl. Math. Letters, 4*(6), 77–80.

Turán, G., & Vatan, F. (1994). On the computation of boolean functions by analog circuits of bounded fan-in. In *Proceedings of the 35th Annual IEEE Symposium* *on Foundations of Computer Science* (pp. 553–564). New York: IEEE Computer Science Press.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications, 16*, 264–280.

von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components. In C. E. Shannon & J. E. McCarthy (Eds.), *Automata studies* (pp. 329–378). Annals of Mathematics Studies 34. Princeton, NJ: Princeton University Press.

Received October 3, 1996; accepted June 3, 1997.