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

Wolfgang Maass
Institute for Theoretical Computer Science
Technische Universität Graz*

Pekka Orponen
Department of Mathematics
University of Jyväskylä†

September 23, 1996

Abstract

We introduce a model for analog noise in analog computation with discrete time that is flexible enough to cover the most important concrete cases, such as noisy analog neural nets and networks of spiking neurons. 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 obviously a serious issue in analog computation. In particular one has to be concerned about the effect of omnipresent smaller (and occasionally larger) perturbations of the values of intermediate analog variables (for example given by some Gaussian distribution). However there exists no formal model which allows us to address this issue in an adequate manner. In addition there exists no noise-model that can also be applied to computations in networks of noisy spiking neurons. Because of the lack of an adequate mathematical model for analog noise it has so far been impossible to prove rigorous bounds for the computational power of discrete-time systems with analog noise.

Previously [Siegelmann, 1994] had investigated the effect of "digital noise" in analog neural nets, where instead of perturbations of intermediate analog values only randomly occuring *total failures* of gates and wires are allowed (similar as in the quite large literature on noisy *digital* circuits [von Neumann, 1956], [Pippenger, 1989]). More recently [Casey, 1996] has investigated a model for analog noise that allows perturbations of intermediate analog variables up to some $\eta > 0$. However, this model does not even cover the case of a Gaussian distribution of noise where also larger deviations than η may occur, although with low probability. Furthermore within his framework one can only analyze those computations which can give their binary output (accept/reject) in spite of noise with *perfect* reliability. Obviously this covers only a rather special class of computations. In addition the model from [Casey, 1996]

*Klosterwiesgasse 32/2, A–8010 Graz, Austria. E-mail: maass@igi.tu-graz.ac.at.

†P. O. Box 35, FIN–40351 Jyväskylä, Finland. E-mail: orponen@math.jyu.fi. Part of this work was done while this author was at the University of Helsinki, and during visits to the Technische Universität Graz and the University of Chile in Santiago.

is not flexible enough to be applicable to networks of noisy spiking neurons, where apart from jitter in the firing times it also occurs that neurons fire "spontaneously", or that neurons fail to fire with a certain probability. In order to model such noise one also has to allow that a discrete state (*"not-firing"*) is randomly replaced by an analog value (a firing time), and vice versa.

As the general setup for our noise model, which is intended to cover all these and in addition any other reasonable concrete situations, let us consider some "analog" system M with d components that each have some bounded real-valued state. The total state space of the system is then some bounded set $\Omega \subseteq \mathbf{R}^d$. In the case of, say, d independent analog neurons, each with a state from the interval $[-1, 1]$, the state space would be simply $\Omega = [-1, 1]^d$, but we wish to allow also for possible correlations between the components, and "forbidden regions" in the state space, e.g. around the origin. Also, we want to cover systems with discrete and hybrid continuous-discrete state spaces, such as networks of spiking neurons [Maass, 1996a] or analog machines with binary registers. The only properties of the state space that are essential to our results are that it is finite-dimensional and *bounded*, i.e. that the components have some hard limits beyond which their response cannot go.

In a noise-free setting, the system M would move in a single time step from a state $p \in \Omega$ to its next state $p^+ \in \Omega$ according to some deterministic rule. (We shall discuss the precise definitions in Section 2.) In the more realistic *noisy* version we consider, the next state q is given only as some *probability distribution* over Ω, with a density function $\pi(p, q)$. This transition density is determined as a function of the deterministic next state p^+, and the *noise process* affecting the system; more precisely, $\pi(p, q) = z(p^+, q)$, where z is the noise density function. In typical situations, this distribution would be centered around p^+, as when z is e.g. some Gaussian density function; however we shall make only the much weaker assumption that the noise process is "piecewise uniformly continuous". (For precise definitions, see Section 2. An issue requiring some care here is that since the state space Ω is bounded, any unbounded, e.g. Gaussian density would have to be "clipped" at the boundaries of Ω, so that any noise leading a component beyond its response limit is "rounded" to that limit value. This causes a technical problem in the context of density functions, since a single state on the boundary of Ω may be assumed with probability > 0 .) All the standard continuous noise processes (and their clipped versions) satisfy our assumptions, as well as noise processes that are smooth over continuous parts of Ω, and defined in an arbitrary manner over the possible discrete parts. We even allow the noise process to depend on the current state of the system in a rather arbitrary fashion.

The first result we prove in this paper (Section 3) is that analog systems of the type discussed above, and affected by noise conforming to our general noise model, have at best the power of finite automata, even if they are allowed to give their binary output with low reliability. This is in stark contrast to e.g. the results in [Siegelmann, Sontag, 1995], where noise-free analog neural nets are shown to be capable of universal computation, and the results in [Maass, 1996a], where the same is proved of the spiking neural network model without noise.

To state the result more precisely, let us assume that our systems are provided with some input-output mechanism, and say that a set L of input sequences is *recognized* by a system M with *reliability* $\rho \geq 0$, if for any input sequence x, the output of M agrees with whether $x \in L$ with probability at least $1/2 + \rho$. We prove that if a set L is recognized with any reliability $\rho > 0$ by a finite-dimensional, bounded system M affected by a piecewise uniformly continuous noise process, then L is regular, i.e. it can also be recognized by some finite automaton.

Conversely, we verify in Section 4 that if we only consider noise processes that are bounded by some constant $\eta < 1/2$ (i.e. for any state p, $z(p,q) > 0$ only for states q such that $|q_i - p_i| \leq \eta$ holds for all i), then any regular set L can be recognized with perfect ($\rho = 1/2$) reliability by a recurrent analog neural net of the type discussed in [Siegelmann, Sontag, 1995], even in the presence of analog noise.

In Section 5 we prove 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, e.g. 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, 1995] 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. We show that this anomaly disappears as soon as even an arbitrarily small amount of analog noise is present.

More specific hardware-oriented models for analog noise in analog neural nets have been discussed in [Phatak, Koren 1995]. 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]).

Finally we would like to point out that model for analog noise can also be applied to completely different types of models for discrete time analog computation, such as arithmetical circuits [Turán, Vatan, 1994], the random access machine (RAM) with analog inputs, the parallel random access machine (PRAM) with analog inputs, various computational discrete-time dynamical systems ([Moore, 1990], [Asarin, Maler, 1994], [Orponen, Matamala, 1996]), and (with some minor adjustments) also the BSS model [Blum, Shub, Smale, 1989], [Koiran et al., 1994]. This article provides for each of these quite important special cases a suitable framework for modeling analog noise, as well as significant results regarding the effect of analog noise on their computational power.

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 5-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 shall give a more precise definition of the dynamics, including the effects of noise, later.)

For instance, the *recurrent analog neural net* model of [Siegelmann, Sontag, 1995] (also known as the *"Brain State in a Box"* model of [Anderson et al., 1977]) 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 $s : \Omega \times \Sigma \to \Omega$ 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* $\sigma : \mathbf{R} \to [-1,1]$. For any $p \in \Omega$ and $a \in \Sigma$, we define $s(p,a) = p^+$, where for each $i = 1,\dots,d$,

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

Both [Anderson et al., 1977] and [Siegelmann, Sontag, 1995] use the saturated-linear sigmoid activation function

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

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

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

the latter choice yielding the model of *recurrent threshold logic networks*. The initial state in each of these models may be chosen as $p^0 = (-1,\dots,-1)$, 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 also the componentwise extension of the chosen activation function to state vectors, so that for any $p \in \Omega$, $\sigma(p) := (\sigma(p_1),\dots,\sigma(p_d))$. 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 l-th 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, 1996]) 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 $[0,T)$, $[T,2T)$, ... , 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, e.g. $l = 15$ 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], i.e., 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 $p \in \Omega$, i.e. $\mu(B) = 0$ 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, i.e. 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 $\pi_x(p^0,q)$, i.e. 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 $x \in \Sigma_0^*$:

$$x \in L \quad \Leftrightarrow \quad \int_F \pi_{xu}(q)\ d\mu \geq \frac{1}{2} + \rho \ \text{ for some } u \in \{\sqcup\}^*$$

$$x \notin L \quad \Leftrightarrow \quad \int_F \pi_{xz}(q)\, d\mu \leq \frac{1}{2} - \rho \ \text{ for all } z \in \{\sqcup\}^*.$$

This covers also 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 if one assumes 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 $\mu_1 + \ldots + \mu_k$, 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" (i.e. 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 via 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 for analog computation without noise (such as recurrent neural nets [Siegelmann, Sontag, 1995] and networks of spiking neurons [Maass, 1996a]) that they can simulate a universal Turing machine, and hence have immense computational power. It has long been conjectured that their computational power collapses 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. It turns out that one has to argue quite differently than in the analysis of *digital* noise (e.g. gate failures) [von Neumann, 1956, Pippenger, 1989] and also completely differently than in the case of noisy analog computations with perfect reliability (i.e. $\rho = \frac{1}{2}$) which had been considered in [Casey, 1996]. A closer relative is Rabin's [Rabin, 1963, Paz, 1971] proof showing that probabilistic finite automata with "isolated cutpoints" are equivalent to deterministic finite automata; however in Rabin's case the finiteness of the state space simplifies the setup considerably.

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 independently 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 uniformly continuous* if for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every $r \in \Omega$, and for all $p, q \in \Omega_j$, $j = 1,\ldots,k$:

$$\| p - q \| \le \delta \quad \text{implies} \quad |\pi(r,p) - \pi(r,q)| \le \varepsilon. \tag{1}$$

If $z(\cdot,\cdot)$ satisfies this condition, we say that the resulting noise process Z is *piecewise uniformly continuous.*

Our preceding discussions suggest that the analog noise process Z is in all practically relevant cases piecewise uniformly continuous. Note that because the state space Ω is bounded, any piecewise uniformly continuous density function on Ω has a bounded range.

To formulate our first 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 only accept 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, independently of the length of the input string.

Theorem 3.1 *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 uniformly continuous 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 $\int_{r\in\Omega} |\pi_x(r) - \pi_y(r)| \; d\mu \le \rho$, 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 $\int_F \pi_{xwu}(q) \; d\mu \ge \frac{1}{2} + \rho$ and $\int_F \pi_{ywu}(q) \; d\mu \le \frac{1}{2} - \rho$. This yields the contradiction

$$2\rho \;\le\; \left| \int_{q\in F} \pi_{xwu}(q) \; d\mu - \int_{q\in F} \pi_{ywu}(q) \; d\mu \right|$$

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

Thus we have shown that $\int_{r\in\Omega} |\pi_x(r) - \pi_y(r)|\; d\mu \le \rho$ 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 (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 $\| p - q \| \le \delta$:

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

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 $\int_{r\in\Omega} |\pi_x(r) - \pi_y(r)|\; d\mu \le \rho$, 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 gridpoint closest to r, then $|\pi_x(r) - \pi_x(t_r)| \le \rho/4\mu(\Omega)$. (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 (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 gridpoint $t \in G$ 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 $r \in \Omega_j \subseteq \Omega$, $j = 1, \ldots, k$,

$$|\pi_x(r) - \pi_y(r)| \le |\pi_x(r) - \pi_x(t_r)| + |\pi_x(t_r) - \pi_y(t_r)| + |\pi_y(t_r) - \pi_y(r)| \le \rho/\mu(\Omega),$$

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

Remark 3.2 *In stark contrast to the results of [Siegelmann, Sontag, 1995] 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.*

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 $\mathbf{R}^d$, i.e. 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 (i.e. $\rho = \frac{1}{2}$) 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 this proof is simply to first 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 vs. delay tradeoff 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 $s(p,a) = \text{sgn}(Wp + h + ac)$, 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, i.e., 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 4.1 *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 $\mathcal{T}$ (i.e., $w_{\max} = \max_i \sum_j |w_{ij}|$). 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 to always operate close to saturation (i.e. 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 an d-unit threshold logic network $\mathcal{T}$ with transition function $p^+ = s(p,a) = \text{sgn}(Wp + h + ac)$, 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:

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

By our choice of the value u_δ, we are thus again assured 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 4.2 *For every regular language $L \subseteq \{-1,1\}^*$ there is a constant $\eta > 0$ such that L can be recognized with perfect reliability (i.e. $\rho = \frac{1}{2}$) 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 e.g. [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 present 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 both 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 4.1 with $\theta = 1$, $w_{\max} = 2m+1$, we may choose $\eta = 1/(2m+1)$. ∎

The next corollary shows that we can increase the noise tolerance of a network by slowing down 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 $a_1^\tau \ldots a_k^\tau$ (i.e. each input symbol a_i is repeated τ times before the next one is presented).

Corollary 4.3 *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 (i.e. $\rho = \frac{1}{2}$) 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 4.2 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 4.4 *One can obtain different noise-tolerance vs. delay tradoffs 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, 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 4.3 achieve a noise-tolerance bound of $\eta = O(1/\sqrt{m \log m})$ with delay $\tau = 4$.

Remark 4.5 *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 e.g. [Siegelmann, Sontag, 1995]), the η bound in Corollary 4.3 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. This novel type of upper bound depends apart from the analog noise only on 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" (i.e. weights etc.). It will become obvious from the proof that this 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.

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 et al., 1989], [Maass, 1995]). It should be noted that this does not only hold for the "classical" 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, Sontag, 1995] 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 will talk in this section also about 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

$$\mathcal{F} := \{f : \mathbf{R}^{n+1} \to \{0,1\} : \exists\, g \in \mathcal{G}(f(\underline{x}, y) = 1 \text{ if } g(\underline{x}) \geq y \text{ and } f(\underline{x}, y) \;=\; 0 \text{ if } g(\underline{x}) < y)\}$$

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 uniformly continuous 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 5.1. 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 uniformly continuous 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

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

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 (2) may come from two different systems in $\mathcal{M}$ (for example from two different weight-assignments to a recurrent analog neural net).

In Theorem 5.1 we consider an even more general set-up 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 5.3.).

We assume in Theorem 5.1 that $\mathcal{G}$ is an arbitrary class of piecewise uniformly continuous density kernels $\pi : \mathbf{R}^n \times \Omega \to [0,B]$ with uniformly bounded moduli of continuity (as in condition (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 3.1) superimpose on the space Ω a finite grid G, such that for any $\pi, \tilde{\pi} \in \mathcal{G}$ and $x, \tilde{x} \in \mathbf{R}^n$:

$$\begin{array}{c} |\pi(x,q) - \tilde{\pi}(\tilde{x},q)|\ \leq \rho/5\mu(\Omega) \text{ for all } q \in G \text{ implies that} \\ \int_{q\in\Omega} |\pi(x,q) - \tilde{\pi}(\tilde{x},q)|\ d\mu < \rho/2. \end{array}$$

The size $|G|$ of the grid (i.e. the number of gridpoints) 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 5.1 *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

$$\begin{array}{c} \int_{p\in F}\int_{q\in\Omega} \pi_A(x,q) \cdot \omega_A(q,p)\ d\mu\ d\mu' \text{ has a value } \geq \frac{1}{2}+\rho \text{ if } x \in A, \\ \text{and else a value } \leq \frac{1}{2}-\rho\ . \end{array}$$

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 $b := \lfloor B/\gamma \rfloor + 1$. One should 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 et al., 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

$$|\mathcal{G}^*| \leq 2 \cdot (|S| \cdot |G| \cdot b^2)^{\beta(|S| \cdot |G|, b, \Delta)}, \tag{3}$$

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 $\mathcal{G}_S^\gamma$ which is induced by $\mathcal{G}^*$ we can now show that the cardinality $|S|$ of the shattered set S can be bounded through the inequality

$$2^{|S|} \leq 2 \cdot (|S| \cdot |G| \cdot b^2)^{\beta(|S| \cdot |G|, b, \Delta)} \cdot 2^{b^{|G|}}. \tag{4}$$

It is obvious that this inequality yields an upper bound for $|S|$ which 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 $\hat{\omega} : \{0, \ldots, b-1\}^G \to \{0,1\}$ which is induced by ω through the following definition:

$$\begin{gathered}
\hat{\omega}(\hat{\pi}) \text{ has value } 0 \text{ for } \hat{\pi} \in \{0, \ldots, b-1\}^G \\
\text{if there exist some } \pi \in \mathcal{G} \text{ and } x \in S \text{ with} \\
|\hat{\pi}(q) - \pi^\gamma(x,q)| \leq 1 \text{ for all } q \in G \text{ and} \\
\int_{p \in F} \int_{q \in \Omega} \pi(x,q) \cdot \omega(q,p)\ d\mu\ d\mu' \leq \tfrac{1}{2} - \rho. \\
\text{Else we set } \hat{\omega}(\hat{\pi}) = 1.
\end{gathered}$$

Since we have the upper bound (3) on the size of the cover $\mathcal{G}^*$, and there exist at most $2^{b^{|G|}}$ different functions $\hat{\omega}$, it suffices for proving (4) 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 $|\tilde{\pi}^\gamma(x,q) - \pi^\gamma_{A_1}(x,q)| \leq 1$ and $|\tilde{\pi}^\gamma(x,q) - \pi^\gamma_{A_2}(x,q)| \leq 1$ 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 then $\hat{\omega}_{A_1} \neq \hat{\omega}_{A_2}$.

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 $\hat{\pi} : G \to \{0, \ldots, b-1\}$ 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 $|\tilde{\pi}^\gamma(x_0,q) - \pi^\gamma_{A_2}(x_0,q)| \leq 1$ for all $q \in G$ and

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

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

$$\int_{p\in F} \int_{q\in\Omega} \pi(x_1,q) \cdot \omega_{A_1}(q,p)\, d\mu\, d\mu' \leq \frac{1}{2} - \rho \quad \text{and} \tag{5}$$

$$|\hat{\pi}(q) - \pi^\gamma(x_1,q)| \leq 1 \text{ for all } q \in G.$$

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

$$\int_{q\in\Omega} |\tilde{\pi}(x_0,q) - \pi(x_1,q)|\, d\mu \quad < \quad \rho/2. \tag{6}$$

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

$$\int_{q\in\Omega} |\tilde{\pi}(x_0,q) - \pi_{A_1}(x_0,q)|\, d\mu \quad < \quad \rho/2\ . \tag{7}$$

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

$$\int_{p\in F} \int_{q\in\Omega} \pi_{A_1}(x_0,q) \cdot \omega_{A_1}(q,p)\, d\mu\, d\mu' \quad \geq \quad \frac{1}{2} + \rho. \tag{8}$$

The inequalities (6) and (7) 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 (5) and (8) the contradiction

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

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 5.1 is now complete. ∎

Remark 5.2 *It follows from [Alon et al., 1993] that instead of a finite upper bound for the* <u>*pseudo-dimension*</u> *of* $\mathcal{G}$ *it suffices for Theorem 5.1 to assume a finite upper bound for the* <u>*γ-dimension*</u> *P_γ-dim$(\mathcal{G})$ of $\mathcal{G}$ for $\gamma = \rho/20\mu(\Omega)$.*

Corollary 5.3 *There exists a finite upper bound for the VC-dimension of layered feedforward sigmoidal neural nets and feedforward networks of spiking neurons with piecewise uniformly continuous analog noise (for arbitrary real-valued inputs, Boolean output computed with some arbitrary reliability $\rho > 0$, and arbitrary real-valued "programmable parameters") which does* <u>*not*</u> *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 uniformly continuous 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 5.1 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, Macintyre, 1995] 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, 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 5.1 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 which determine other aspects of the functional form of response functions (i.e. postsynaptic potentials) and threshold functions. The pseudo-dimension of the class $\mathcal{G}$ which arises when one applies (as described in Section 2) the here considered framework 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 5.4 *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") which 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 5.3, 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 previously been used for the investigation of noisy computations.

Finally we have established significant new results regarding general limitations for the computational power and VC-dimension of noisy computational systems. These results provide in particular new bounds for the computational power and VC-dimension of analog neural nets and networks of spiking neurons in the presence of analog noise.

Acknowledgement

We would like to thank Peter Auer for helpful conversations.

References

[Alon et al., 1993] N. Alon, N. Cesa-Bianchi, S. Ben-David, D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability. *Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science*, 292–301. IEEE Computer Science Press, New York, NY, 1993.

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

[Anderson et al., 1977] J. A. Anderson, J. W. Silverstein, S. A. Ritz, R. S. Jones. Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review 84* (1977), 413–451. Reprinted in [Anderson, Rosenfeld, 1988], pp. 287–325.

[Anderson, Rosenfeld, 1988] J. A. Anderson and E. Rosenfeld (eds.), *Neurocomputing: Foundations of Research.* The MIT Press, Cambridge, MA, 1988.

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

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

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

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

[Feller, 1971] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. 2, Second Edition.* J. Wiley & Sons, New York, NY, 1971.

[Gerstner, van Hemmen, 1994] W. Gerstner, J. L. van Hemmen, How to describe neuronal activity: spikes, rates or assemblies? *Advances in Neural Information Processing Systems 6*, 463–470. Morgan Kaufmann, San Mateo, CA, 1994.

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

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

[Hopcroft, Ullman, 1979] J. E. Hopcroft, J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, Reading, MA, 1979.

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

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

[Karpinski, Macintyre, 1995] M. Karpinski, A. Macintyre, Polynomial bounds for VC-dimension of sigmoidal and general Pfaffian neural networks. *J. of Computer and System Sciences*, to appear.

[Kifer, 1988] Y. Kifer, *Random Perturbations of Dynamical Systems.* Birkhäuser, Boston, MA, 1988.

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

[Maass, 1995] W. Maass, Vapnik-Chervonenkis dimension of neural nets. *The Handbook of Brain Theory and Neural Networks* (M. A. Arbib, ed.), 1000–1003. The MIT Press, Cambridge, MA, 1995.

[Maass, 1996] W. Maass, On the computational power of noisy spiking neurons. *Advances in Neural Information Processing Systems 8*, 211–217. The MIT Press, Cambridge, MA.

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

[Minsky, 1972] M. L. Minsky, *Computation: Finite and Infinite Machines.* Prentice-Hall, Englewood Cliffs, NJ, 1972.

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

[Orponen, Matamala, 1996] P. Orponen, M. Matamala, Universal computation by finite two-dimensional coupled map lattices. Manuscript, *Proc. Workshop on Physics and Computation PhysComp'96*, to appear.

[Paz, 1971] A. Paz, *Introduction to Probabilistic Automata.* Academic Press, New York, NY, 1971.

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

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

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

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

[Siegelmann, Sontag, 1995] H. T. Siegelmann, E. D. Sontag, On the computational power of neural nets. *J. Comput. System Sciences 50* (1995), 132–150.

[Turán, Vatan, 1994] G. Turán, F. Vatan, On the computation of Boolean functions by analog circuits of bounded fan-in. *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science*, 553–564. IEEE Computer Science Press, New York, NY, 1994.

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

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