Shot-Noise-Limited Performance of Optical Neural Networks

Majeed M. Hayat, *Member, IEEE*, Bahaa E. A. Saleh, *Fellow, IEEE*, and John A. Gubner, *Member, IEEE*

***Abstract*— The performance of neural networks for which weights and signals are modeled by shot-noise processes is considered. Examples of such networks are optical neural networks and biological systems. We develop a theory that facilitates the computation of the average probability of error in binary-input/binary-output multistage and recurrent networks. We express the probability of error in terms of two key parameters: the computing-noise parameter and the weight-recording-noise parameter. The former is the average number of particles per clock cycle per signal and it represents noise due the particle nature of the signal. The latter represents noise in the weight-recording process and is the average number of particles per weight. For a fixed computing-noise parameter, the probability of error decreases with the increase in the recording-noise parameter and saturates at a level limited by the computing-noise parameter. A similar behavior is observed when the role of the two parameters is interchanged. As both parameters increase, the probability of error decreases to zero exponentially fast at a rate that is determined using large deviations. We show that the performance can be optimized by a selective choice of the nonlinearity threshold levels. For recurrent networks, as the number of iterations increases, the probability of error increases initially and then saturates at a level determined by the stationary distribution of a Markov chain.**

I. Introduction

NOISE plays an important role in determining the performance of neural networks. Noise takes the form of fluctuations of the signals involved in the computation, and uncertainty of the weights and other parameters of the network. This inaccuracy accumulates as the signals propagate through multistage or recurrent networks, so that the actual final output may become different from the desired output, resulting in errors. Previous studies that have been concerned with the sensitivity of neural networks to signal fluctuations and weight uncertainty employed various Gaussian and other approximations [16], [2], [6], [17], [8]. Such Gaussian and signal-independent noise models are inadequate for optical and biological networks in which the noise described by shot-noise processes which arise as a result of the underlying particle nature of the signals, e.g., photons in optical beams or neural spikes in biological systems [3], [11]. A shot-noise process is a filtered Poisson point process whose rate may also be random.

While this particle noise has particularly deleterious effects at low particle fluxes [3], [12], which are associated with weak signals, its signal-dependent nature has an important effect on the errors, even if strong signals are used, if very low error rates are to be accomplished. In this paper, we provide an analysis of the performance of such networks in an attempt to determine how strong the signals must be to achieve desired levels of accuracy. This is important in networks with very large number of inputs since the total signal power is constrained. In this analysis, we ignore other sources of noise and uncertainty and focus on the fundamental limiting factor, which is the underlying particle nature of the noise. Although the paper is presented in the context of optical neural networks in which there is currently a great deal of interest, the results apply to other shot-noise limited networks. Our aim is to determine the fundamental limits on optical networks, set by the quantum nature of light, which can be quite restrictive if high data throughputs are to be accomplished.

The basic unit of a neural network involves incoming signals which are multiplied by weights, and then added and thresholded to produce the outputs. The signals are described by shot-noise processes. In addition, the weights themselves are random variables resulting from sampled shot-noise processes. In an all-optical system, for example, the weights are recorded by optical beams, each described by a shot-noise process [13], [14]. In biological systems, the weights are dynamically altered by signals originating from a nerve-spike train and are also modeled as shot-noise processes. The noise in the signals is referred to as computing noise, and that in the weights is called the weight-recording noise. The errors generated by these two noise sources are primarily governed by the flux of particles underlying the computing and recording signals (e.g., average photon flux or nerve spike rate).

Modeling signals and weights by shot noise processes, we provide a probabilistic analysis that determines the probability of error in neural networks of simple architectures. The analysis is tailored for binary-input/binary-output networks with threshold (hardlimiter) nonlinearities. Typical examples of these networks are rule forming, global classifier, and Hopfield networks [10]. The weight elements are all assumed to be nonnegative. This assumption was shown to be desirable in some networks because it leads to superior performance [6], and it simplifies the analysis. Nonetheless, our analysis can be easily extended to handle two-channel systems with a concomitant subtraction step [5].

Manuscript received May 17, 1994; revised January 5, 1995 and October 25, 1995. Presented in part at the Optical Society of America Annual Meeting, Dallas, TX, October 1994. This work was supported by the Office of Naval Research Grant N00014–94–1–0366, the Office of Naval Research Grant N00014–94–1–0366, and by the Air Force Office of Scientific Research Grant F49620–92–J–0305.

M. M. Hayat and J. A. Gubner are with the Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI 53706–1691 USA.

B. E. A. Saleh is with the Department of Electrical, Computer and Systems Engineering, Boston University, Boston, MA 02215-2407 USA.

Publisher Item Identifier S 1045-9227(96)02882-2.

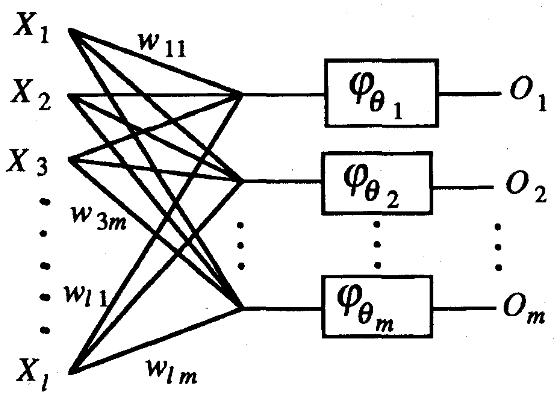


Fig. 1. An l-input/m-output single-layer neural network with weight matrix $\mathbf{W}$ and threshold nonlinearity vector function Φ_Θ.

II. Model

Consider the single-layer (Adaline) network shown in Fig. 1. This can be thought of as the kth layer of a feedforward (Madaline) network for example. Let $\mathbf{X} \in \{0,1\}^l$ denote an l-dimensional random binary input with probability mass function $f(\mathbf{x}) \stackrel{\triangle}{=} \mathrm{P}\{\mathbf{X} = \mathbf{x}\}$. Let $\mathbf{W}$ be an $m \times l$ matrix with nonnegative, real-valued entries w_{ij} corresponding to the network weights. The jth component of the output $\mathbf{O}$ of the network is given by $O_j = \varphi_{\theta_j}(W_j\mathbf{X})$, where W_j denotes the jth row of $\mathbf{W}$, and the function φ_{θ_j} is a $\{0,1\}$-valued saturation point nonlinearity with a nonnegative threshold constant θ_j

$$\varphi_{\theta_j}(u) = \begin{cases} 1, & u > \theta_j, \\ 0, & u \le \theta_j. \end{cases}$$

More compactly

$$\mathbf{O} = \Phi_\Theta(\mathbf{W}\mathbf{X}) \tag{1}$$

where for each $\mathbf{u} = [u_1, \cdots, u_m]'$ in $\mathbb{R}^m$, $\Phi_\Theta(\mathbf{u}) = [\varphi_{\theta_1}(u_1), \cdots, \varphi_{\theta_m}(u_m)]'$, and the prime denotes the transpose of a vector. (We also allow nonlinearities of the form $1 - \varphi_{\theta_j}$.)

We assume that each component X_i of $\mathbf{X}$, $(i = 1, \ldots, l)$, is represented by a shot-noise process $\tilde{X}_i$ with an underlying Poisson process whose intensity is proportional to X_i and an appropriately scaled filter so that $\mathrm{E}[\tilde{X}_i]$ is proportional to X_i. Similarly, we assume that each weight element w_{ij}, $(i = 1, \ldots, m, j = 1, \ldots, l)$, is represented by a shot-noise process A_{ij} with an underlying Poisson process with an intensity proportional to w_{ij} and an appropriately scaled filter so that $\mathrm{E}[A_{ij}]$ is proportional to w_{ij}. The role of A_{ij} in the multiplication operation is to modify the intensity of the shot-noise process $\tilde{X}_i$ by a multiplicative factor A_{ij}. These modified shot-noise processes are then added to produce the signal Y_j which then becomes the input to a threshold nonlinearity. We will show that the above model is applicable to optical neural networks. We thereafter present the results of this paper in the context of optical neural networks.

In an optical implementation of the network in Fig. 1, each component X_i of $\mathbf{X}$, $(i = 1, \cdots, l)$, is set up to generate an optical beam of intensity λX_i, where the factor λ controls the optical intensity of each beam. The array of beams is transmitted through a transparency with an array of transmittances corresponding to a matrix $\tilde{\mathbf{A}}$. Each entry $\tilde{A}_{ij}(x,y)$ of $\tilde{\mathbf{A}}$ corresponds to a weight w_{ij}, and it is prepared as follows [14]: A blank transparency is exposed to a beam of intensity μw_{ij} for a duration τ_r seconds to produce the shot-noise

$$\tilde{A}_{ij}(x,y) = \mu^{-1}\tau_r^{-1}\sum_k \tilde{g}(x - x_k, y - y_k) \tag{2}$$

arising from a spatial filter $\mu^{-1}\tau_r^{-1}\tilde{g}(\cdot,\cdot)$ ($\tilde{g}$ is the point spread function of the transparency) and an underlying spatial Poisson process of density $\mu w_{ij}\tau_r$ (per unit area). (The x, y variables for each $\tilde{A}_{ij}$ are measured from the center of the (i,j)th transparency.) We assume that $\tilde{g}(\cdot,\cdot)$ is zero outside the disc centered at the origin and of area $\Delta_{\tilde{g}}$. It is also assumed that $\tilde{\mathbf{A}}$ and $\mathbf{X}$ are mutually independent and that the entries of $\tilde{\mathbf{A}}$ are also mutually independent. A simple calculation shows that $\mathrm{E}[\tilde{A}_{ij}(x,y)] = w_{ij}\iint_{\mathbb{R}^2}\tilde{g}(x,y)\,dxdy$ and hence because of the normalization by $\mu^{-1}\tau_r^{-1}$ in (2), the transmittance $\tilde{A}_{ij}$ is proportional to the desired weight w_{ij} in the mean. The parameter μ, as we will see, determines the accuracy level of the recorded transparency.

The ith beam of the m-dimensional array of beams generated from the optical multiplication has an optical intensity $\lambda\tilde{A}_i(x,y)\mathbf{X}$ where $\tilde{A}_i(x,y)$ is the ith row of $\tilde{\mathbf{A}}$. Note that for each beam, the intensity varies from point to point within the beam cross section. The array of beams is then detected by an array of photodetectors. Each photodetector responds to the integrated optical intensity over its active area consisting of a disc $\mathcal{D}_d$ of area Δ_d. The output of ith photodetector is a temporal shot noise process $\tilde{Y}_i(t)$ generated by a causal filter $h(\cdot)$ (of duration τ_c, where τ_c denotes the computing time) and an underlying doubly stochastic Poisson process $\{t_k\}$ with random rate (per unit time)

$$\Lambda_i = \lambda\iint_{\mathcal{D}_d}\tilde{A}_i(x,y)\mathbf{X}\,dxdy. \tag{3}$$

Namely

$$\tilde{Y}_i(t) = \sum_{0 \le t_k \le \tau_c} h(t - t_k), \quad 0 \le t \le \tau_c.$$

Furthermore, if we define the random variables

$$A_{ij} \stackrel{\triangle}{=} \iint_{\mathcal{D}_d}\tilde{A}_{ij}(x,y)\,dxdy \tag{4}$$

and the function

$$g(x,y) \stackrel{\triangle}{=} \iint_{\mathcal{D}_d}\tilde{g}(x' - x, y' - y)\,dx'dy' \tag{5}$$

then from (2) and (4)

$$A_{ij} = \mu^{-1}\tau_r^{-1}\sum_k g(x_k, y_k) \tag{6}$$

and

$$\Lambda_i = \lambda A_i\mathbf{X} \tag{7}$$

where $A_i = [A_{i1}, A_{i2}, \cdots, A_{il}]$. Hence, each A_{ij} is a shot noise-random variable generated by the filter $\mu^{-1}\tau_r^{-1}g(\cdot,\cdot)$ and an underlying spatial Poisson process with rate $\mu w_{ij}\tau_r$. Note that the function $g(\cdot,\cdot)$ is zero outside the disc $\mathcal{D}$ centered at the origin and of area $\Delta \triangleq \Delta_{\tilde{g}} + \Delta_d + 2\sqrt{\Delta_{\tilde{g}}\Delta_d}$. The parameter Δ represents the spatial resolution of the system.

The output of each detector is then sampled at time τ_c and divided by λ to generate the shot-noise random variable $Y_i \triangleq \tilde{Y}_i(\tau_c)/\lambda$. Note that due to the division by λ, the conditional mean for each Y_i is independent of λ, i.e., $\mathrm{E}[Y_i \,|\, \mathbf{X} = \mathbf{x}] = \gamma\gamma' W_i \mathbf{x}$, where $\gamma = \int_0^{\tau_c} h(t)\,dt$ and $\gamma' = \iint_{\mathcal{D}} g(x,y)\,dxdy$.

In summary, conditioned on $\mathbf{A}$ and $\mathbf{X}$, each Y_i is conditionally a shot-noise random variable generated by a filter $\lambda^{-1}h(\cdot)$ and a temporal Poisson process with intensity Λ_i given by (7), i.e., Y_i is a doubly stochastic shot-noise random variable. Furthermore, since the Y_i's are generated by distinct detectors, they are mutually independent conditional on $\mathbf{X}$.

Finally, each Y_i is then passed through a threshold nonlinearity $\varphi_{\xi_i}(\cdot)$ to yield the ith element of the final binary output vector $\mathbf{Z} = \Phi_{\Xi}(\mathbf{Y})$, where $\Phi_{\Xi}(\mathbf{u}) = [\varphi_{\xi_1}(u_1), \cdots, \varphi_{\xi_m}(u_m)]'$, and $\mathbf{Y} = [Y_1, \cdots, Y_m]'$. Ideally, we would choose $\xi_i = \gamma\gamma'\theta_i$, and we would expect that $\mathbf{Z} = \mathbf{O}$. This may not be the case in general, however, due to the random fluctuation of Y_i around its mean. A more selective choice of ξ_i, as we shall see in Section III-B, may reduce, in the probabilistic sense, the deviation of the optical network from its deterministic counterpart.

III. PERFORMANCE OF SINGLE-LAYER NETWORKS

We are interested in determining the probability of incorrect mapping, namely $\mathrm{P}\{\mathbf{Z} \neq \mathbf{O}\}$, which we denote by $P_e(\lambda, \mu)$, and understand the effect of λ and μ on it. To determine $P_e(\lambda,\mu)$, it is sufficient to compute the conditional probabilities of correct mapping $P_c(\lambda, \mu \,|\, \mathbf{x})$ since

$$\begin{aligned} P_e(\lambda, \mu) &= \mathrm{E}[P_e(\lambda, \mu \,|\, \mathbf{X})] \\ &= 1 - \sum_{\mathbf{x} \in \{0,1\}^l} P_c(\lambda, \mu \,|\, \mathbf{x}) f(\mathbf{x}) \end{aligned}$$

where f is the probability mass function of the random vector $\mathbf{X}$ introduced in Section II. Note that

$$\begin{aligned} P_c(\lambda, \mu \,|\, \mathbf{x}) &= \mathrm{P}\{\mathbf{Y} \in \Phi_{\Xi}^{-1}(\Phi_{\Theta}(\mathbf{W}\mathbf{X})) \,|\, \mathbf{X} = \mathbf{x}\} \\ &= \prod_{i=1}^{m} \mathrm{P}\{Y_i \in \varphi_{\xi_i}^{-1}(\varphi_{\theta_i}(W_i\mathbf{X})) \,|\, \mathbf{X} = \mathbf{x}\} \\ &\triangleq \prod_{i=1}^{m} P_c^i(\lambda, \mu \,|\, \mathbf{x}) \end{aligned}$$

where $P_c^i(\lambda, \mu \,|\, \mathbf{x})$ is the conditional probability of correct mapping of the ith output. Since each φ_{ξ_i} is a threshold nonlinearity with threshold level ξ_i, the set $\varphi_{\xi_i}^{-1}(\varphi_{\theta_i}(W_i\mathbf{x}))$ is either $(-\infty, \xi_i]$ or (ξ_i, ∞) corresponding to $W_i\mathbf{x} \leq \theta_i$ or $W_i\mathbf{x} > \theta_i$, respectively.

To compute $P_c^i(\lambda, \mu \,|\, \mathbf{x})$, we first determine the conditional moment generating functions (MGF's) $Q_{Y_i \,|\, \mathbf{X}, A_i}(s \,|\, \mathbf{x}, \mathbf{a}) \triangleq \mathrm{E}[e^{sY_i} \,|\, \mathbf{X} = \mathbf{x}, A_i = \mathbf{a}]$, $s \in \mathbb{C}$ (the symbol $\mathbb{C}$ denotes the set of complex numbers), and $i = 1, \cdots, m$. It is clear from (7) that once A_i and $\mathbf{X}$ are fixed, the intensity Λ_i will also be fixed (i.e., deterministic), and hence Y_i becomes a shot-noise random variable. Consequently, $Q_{Y_i \,|\, \mathbf{X}, A_i}(s \,|\, \mathbf{x}, \mathbf{a})$ can be computed using the well-known form of a shot noise random variable [15]

$$Q_{Y_i \,|\, \mathbf{X}, A_i}(s \,|\, \mathbf{x}, \mathbf{a}) = \exp\{\lambda \mathbf{a}\mathbf{x}\alpha(s/\lambda)\} \tag{8}$$

where $\mathbf{a}$ is a row vector, $\mathbf{x}$ is a column vector, and

$$\alpha(s) = \int_0^{\tau_c} (e^{sh(t)} - 1)\,dt. \tag{9}$$

We now proceed to remove the conditioning on A_i. Let $Q_{A_i}(\mathbf{s})$ denote the MGF of the random row vector A_i, i.e.,

$$Q_{A_i}(\mathbf{s}) = \mathrm{E}[e^{A_i \mathbf{s}}], \qquad \mathbf{s} \in \mathbb{C}^l. \tag{10}$$

Averaging (8) over all possible A_i we obtain the conditional MGF of Y_i given $\mathbf{X} = \mathbf{x}$

$$Q_{Y_i \,|\, \mathbf{X}}(s \,|\, \mathbf{x}) = Q_{A_i}([\lambda x_1 \alpha(s/\lambda), \cdots, \lambda x_l \alpha(s/\lambda)]') \tag{11}$$

where x_i is the ith coordinate of $\mathbf{x}$. It is clear from (4) and the independence of the $\tilde{A}_i$'s that the components of the random row vector A_i are mutually independent. We also know from the discussion in the preceding section that each element A_{ij} is a shot noise random variable, scaled by $\mu^{-1}\tau_r^{-1}$, resulting from a filter $g(\cdot,\cdot)$ and an underlying Poisson process with rate $\mu w_{ij}\tau_r$. Therefore, (10) takes the special form

$$Q_{A_i}(\mathbf{s}) = \exp\left(\mu\tau_r \sum_{j=1}^{l} w_{ij}\beta(s_j/\mu\tau_r)\right) \tag{12}$$

and hence

$$Q_{Y_i \,|\, \mathbf{X}}(s \,|\, \mathbf{x}) = \exp\left(\mu\tau_r \sum_{j=1}^{l} w_{ij}\beta(\lambda x_j \alpha(s/\lambda)/\mu\tau_r)\right) \tag{13}$$

where

$$\beta(s) = \iint_{\mathcal{D}} (e^{sg(x,y)} - 1)\,dxdy. \tag{14}$$

Using the mean value theorem for integrals (assuming s is real), (13) can be recast in the more informative form

$$\begin{aligned} &Q_{Y_i \,|\, \mathbf{X}}(s \,|\, \mathbf{x}) \\ &\quad = \exp\left(\mu\Delta\tau_r \sum_{j=1}^{l} w_{ij} \right. \\ &\qquad \left. \times \left[\exp\left\{g^*\Delta \frac{\lambda\tau_c}{\mu\Delta\tau_r} x_j (e^{sh^*\tau_c/\lambda\tau_c} - 1)\right\} - 1\right]\right) \end{aligned} \tag{15}$$

where h^* and g^* are intermediate values defined by $h^* = \tau_c^{-1}\int_0^{\tau_c}(e^{sh(t)} - 1)\,dt$ and $g^* = \Delta^{-1}\iint_{\mathcal{D}}(e^{sg(x,y)} - 1)\,dxdy$, respectively. The statistics of Y_i conditioned on $\mathbf{X}$ therefore depend on two parameters: the computing-noise parameter

$$N_c \triangleq \lambda\tau_c \tag{16}$$

and the weight-recording-noise parameter

$$N_r \triangleq \mu\Delta\tau_r. \tag{17}$$

The former is the mean number of photons (which is proportional to optical energy) per computing time in each beam, while the latter is the mean number of photons per recording time per pixel of spatial resolution. These parameters can be cast in the more general context of shot-noise limited systems: N_r is the average number of particles per clock time per signal, and N_c is the average number of particles per weight.

In principle, one can compute conditional probability density functions from conditional MGF's by taking their inverse Laplace transform. This approach, however, is generally difficult to implement numerically. A numerical technique that directly computes the probability $P_c^i(\lambda, \mu \,|\, \mathbf{x})$ from the characteristic function $Q_{Y_i \,|\, \mathbf{X}}(ju \,|\, \mathbf{x})$ has been developed by the authors [7]. We will use this technique in our computations in the examples. The limiting behavior of the average probability of error for large values of μ and λ is studied in the next section.

A. Asymptotic Analysis of the Performance

We start by examining the expression for the conditional MGF of Y_i given in (13). It is easy to see that, $\lambda\alpha(s/\lambda)$ converges to γs as $\lambda \to \infty$. Hence

$$\lim_{\lambda\to\infty} Q_{Y_i \,|\, \mathbf{X}}(s \,|\, \mathbf{x}) = \exp\left(\mu\tau_r \sum_{j=1}^{l} w_{ij}\beta(x_j\gamma s/\mu\tau_r)\right) \triangleq Q_i^{\mu}(s \,|\, \mathbf{x}).$$

Note that we can recognize $Q_i^{\mu}(\cdot \,|\, \mathbf{x})$ as the moment generating function of the sum $D(\mathbf{x}) = \mu^{-1}\tau_r^{-1}x_1\gamma D_1 + \cdots + \mu^{-1}\tau_r^{-1}x_m\gamma D_m$, where each D_j, $j = 1, \cdots, m$, is an independent spatial shot noise process with filter $g(\cdot,\cdot)$ and underlying Poisson process with mean intensity $\mu\tau_r w_{ij}$. Thus, conditioned on $\mathbf{X} = \mathbf{x}$, Y_i converges in distribution to the random variable $D(\mathbf{x})$, and we determine the limiting behavior of the conditional probability of error in the ith output

$$\lim_{\lambda\to\infty} P_e^i(\lambda, \mu \,|\, \mathbf{x}) = \mathrm{P}\{D(\mathbf{x}) \notin \psi_{\xi_i}^{-1}(\varphi_{\theta_i}(W_i\mathbf{x})) \,|\, \mathbf{X} = \mathbf{x}\}. \tag{18}$$

Similarly, we obtain the limiting conditional MGF as $\mu \to \infty$

$$\lim_{\mu\to\infty} Q_{Y_i \,|\, \mathbf{X}}(s \,|\, \mathbf{x}) = \exp\{\lambda\alpha(s/\lambda)\gamma' W_i\mathbf{x}\} \triangleq Q_i^{\lambda}(s \,|\, \mathbf{x}).$$

Observe that $Q_i^{\lambda}(\cdot \,|\, \mathbf{x})$ can be recognized as the MGF of $\hat{D}(\mathbf{x})/\lambda$ where $\hat{D}(\mathbf{x})$ is a shot-noise process resulting from a filter $h(\cdot)$ and an underlying Poisson process of intensity rate $\lambda\gamma' W_i\mathbf{x}$. Hence

$$\lim_{\mu\to\infty} P_e^i(\lambda, \mu \,|\, \mathbf{x}) = \mathrm{P}\{\lambda^{-1}\hat{D}(\mathbf{x}) \notin \varphi_{\xi_i}^{-1}(\varphi_{\theta_i}(W_i\mathbf{X})) \,|\, \mathbf{X} = \mathbf{x}\}. \tag{19}$$

Finally, it is easy to check that

$$\lim_{\lambda,\mu\to\infty} Q_{Y_i \,|\, \mathbf{X}}(s \,|\, \mathbf{x}) = \exp\{\gamma\gamma' W_i\mathbf{x}s\} \triangleq Q_i(s \,|\, \mathbf{x})$$

which is recognized as the conditional MGF of the random variable $\gamma\gamma' W_i\mathbf{X}$, given $\mathbf{X} = \mathbf{x}$. Thus, by fixing $\mathbf{X} = \mathbf{x}$, Y_i converges, as both λ and μ approach ∞, to the constant $\gamma\gamma' W_i\mathbf{x}$ in distribution and hence in probability. Therefore, $\lim_{\lambda,\mu\to\infty} P_e^i(\lambda, \mu \,|\, \mathbf{x})$ exists and it is equal to zero for each fixed $\mathbf{x}$. We now characterize this limit further by providing the rate of convergence. Let $\{\lambda_n\}$ and $\{\mu_n\}$ be two sequences diverging to ∞. We are interested in determining the behavior of

$$P_e^i(\mathbf{x}, n) \triangleq P_e^i(\lambda_n, \mu_n \,|\, \mathbf{x})$$

as $n \to \infty$, for a fixed $\mathbf{x}$. There are three cases to consider:

1) $\lim_{n\to\infty} \frac{\lambda_n}{\mu_n} = k, \qquad 0 < k < \infty$;
2) $\lim_{n\to\infty} \frac{\lambda_n}{\mu_n} = \infty$; and
3) $\lim_{n\to\infty} \frac{\lambda_n}{\mu_n} = 0$.

To address this problem, we appeal to the theory of large deviations. It turns out that the decay is exponential in all cases. We have the following result whose proof uses the Gartner–Ellis theorem [1], [4]. The proof is deferred to the Appendix.

Theorem: For each $i = 1, \cdots, m$, let $\delta_i = \gamma\gamma' W_i\mathbf{x}$, and suppose that $\xi_i > \delta_i$ if and only if $\theta_i > W_i\mathbf{x}$. Then,

For Case 1)

$$r_i(\mathbf{x}) \triangleq \lim_{n\to\infty} \frac{1}{\lambda_n} \log P_e^i(\mathbf{x}, n) = -\rho_{1,i}\xi_i + k^{-1} \sum_{j=1}^{l} w_{ij}\beta(kx_j\alpha(\rho_{1,i})) \tag{20}$$

where $\rho_{1,i}$ is the unique solution to the equation

$$\xi_i = \int_0^{\tau_c} h(t) \exp\{\rho_{1,i}h(t)\}\, dt \sum_{j=1}^{l} w_{ij}x_j \iint_{\mathcal{D}} g(x,y) \exp\{kx_j\alpha(\rho_{1,i})g(x,y)\}\, dx\, dy.$$

For Case 2)

$$r_i(\mathbf{x}) \triangleq \lim_{n\to\infty} \frac{1}{\lambda_n} \log P_e^i(\mathbf{x}, n) = -\rho_{2,i}\xi_i + \gamma'\alpha(\rho_{2,i})W_i\mathbf{x} \tag{21}$$

where $\rho_{2,i}$ is the unique solution to the equation

$$\xi_i = \gamma' W_i\mathbf{x} \int_0^{\tau_c} h(t)e^{\rho_{2,i}h(t)}\, dt.$$

For Case 3)

$$r_i(\mathbf{x}) \triangleq \lim_{n\to\infty} \frac{1}{\mu_n\tau_r} \log P_e^i(\mathbf{x}, n) = -\rho_{3,i}\xi_i + \sum_{j=1}^{l} w_{ij}\beta(x_j\gamma\rho_{3,i}) \tag{22}$$

where $\rho_{3,i}$ is the unique solution to the equation

$$\xi_i = \gamma \sum_{j=1}^{l} w_{ij}x_j \iint_{\mathcal{D}} g(x,y) \times \exp\{x_j\gamma\rho_{3,i}g(x,y)\}\, dx\, dy.$$

The hypothesis in the theorem guarantees that in the limit the optical network becomes equivalent to its deterministic counterpart. Choices of ξ_i that violate the hypothesis should be avoided since this tends to change the task of the network.

Using the theorem, we obtain an expression for the exponential decay rate $r(\mathbf{x})$ of the conditional probability of incorrect mapping of the output vector

$$P_e(\mathbf{x}, n) \overset{\triangle}{=} \mathrm{P}\{\mathbf{Y} \notin \Phi_{\Xi}^{-1}(\Phi_{\Theta}(\mathbf{WX})) \,|\, \mathbf{X} = \mathbf{x}\}$$

as follows:

$$\begin{aligned} r(\mathbf{x}) &\overset{\triangle}{=} \lim_{n\to\infty} \frac{1}{q_n} \log P_e(\mathbf{x}, n) \\ &= \lim_{n\to\infty} \frac{1}{q_n} \log \mathrm{P}\{Y_1 \in \varphi_{\xi_1}^{-1}(\varphi_{\theta_1}(W_1\mathbf{X})), \cdots, \\ &\qquad Y_m \in \varphi_{\xi_m}^{-1}(\varphi_{\theta_m}(W_m\mathbf{X})) \,|\, \mathbf{X} = \mathbf{x}\} \\ &= \lim_{n\to\infty} \frac{1}{q_n} \log \left(1 - \prod_{i=1}^{m}[1 - P_e^i(\mathbf{x}, n)]\right) \\ &= \max_{i=1,\cdots,m} r_i(\mathbf{x}) \end{aligned} \tag{23}$$

where $q_n = \mu_n \tau_r$ for Case 3) and $q_n = \lambda_n$ otherwise.

Finally, the exponential decay rate for the average probability of error $P_e(n) \overset{\triangle}{=} \mathrm{E}[P_e(\mathbf{X}, n)]$ can be computed as follows:

$$\begin{aligned} r &\overset{\triangle}{=} \lim_{n\to\infty} \frac{1}{q_n} \log P_e(n) \\ &= \lim_{n\to\infty} \frac{1}{q_n} \log \mathrm{E}[P_e(\mathbf{X}, n)] \\ &= \lim_{n\to\infty} \frac{1}{q_n} \log \sum_{\mathbf{x}} P_e(\mathbf{x}, n) f(\mathbf{x}) \\ &= \max_{\mathbf{x}\in\{0,1\}^l} r(\mathbf{x}). \end{aligned} \tag{24}$$

Example 1: Consider the optical implementation of the neural network of Fig. 1 with $m = 1$. Two cases are studied: In the first we take $l = 2$ and $\mathbf{W} = [0.5, 0.5]$; in the second $l = 100$ and $\mathbf{W} = [0.01, \cdots, 0.01]$. The threshold $\theta_1 = 0.75$ in both cases. It is assumed that all inputs are equiprobable. The temporal and spatial shot-noise filters are chosen as follows:

$$h(t) = \begin{cases} \tau_c^{-1}, & 0 \le t \le \tau_c, \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

and

$$g(x, y) = \begin{cases} \Delta^{-1}, & x^2 + y^2 \le \Delta/\pi, \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

(In particular, the above form of g is obtained if $\tilde{g}$ is a delta function; this corresponds physically to the case when the point spread function of the transparency is much narrower than the beam width and the detector's active area.) For this example, $h^*\tau_c$ and $g^*\Delta$ appearing in the conditional MGF (15) are both equal to one, and we obtain

$$\begin{aligned} &Q_{Y_1 \,|\, \mathbf{X}}(s, \mathbf{x}) \\ &\quad = \exp\left(\frac{1}{l} N_r \sum_{j=1}^{l} \left[\exp\left(N_c N_r^{-1} x_j (e^{s/N_c}) - 1\right) - 1\right]\right) \end{aligned} \tag{27}$$

with $l = 2$ and $l = 100$, corresponding to the first case and the second case, respectively. For example, if we take $\tau_c = 10^{-9}$ s (corresponding to the response time of a fast photodetector) and use a 100 pW beam of wavelength 1 μm, then $N_c = 503$ photons. If the detector's active surface matches the beam, and if the recording time τ_r is also 10^{-9} s, then $N_r = 503$ photons. Fig. 2(a) depicts the dependence of $P_e(\lambda, \mu)$ on the computing-noise parameter N_c for fixed values of the weight-recording-noise parameter N_r. The curves labeled with diamonds correspond to the case $l = 100$. As N_c increases, $P_e(\lambda, \mu)$ approaches the constant $\mathrm{E}[\lim_{\mu\to\infty} P_e^i(\lambda, \mu \,|\, \mathbf{X})]$, where the quantity inside the expectation is given by (18). A similar plot is obtained if the roles of N_c and N_r are reversed. Finally, $P_e(\lambda, \mu)$ converges to zero exponentially fast as $N_c = N_r$ approach ∞ [see Fig. 2(b)]. The exponential rate (with respect to N_c), for the case $l = 2$, is computed from (20), (23), and (24): $r = -0.5[1.5\rho + 2 - 2\exp(e^\rho - 1)]$, where ρ is the solution to the equation $0.75 = \exp(\rho - 1 + e^\rho)$. These equations yield $r = -0.0175$, which is in agreement with Fig. 2(b). The important conclusion extracted from this example is that to achieve a particular accuracy (i.e., for a fixed average probability of error), there is trade-off, on the one hand, between spatial resolution Δ, recording optical power μ, and recording time τ_r; and on the other hand, between computing speed τ_c and processing optical power λ.

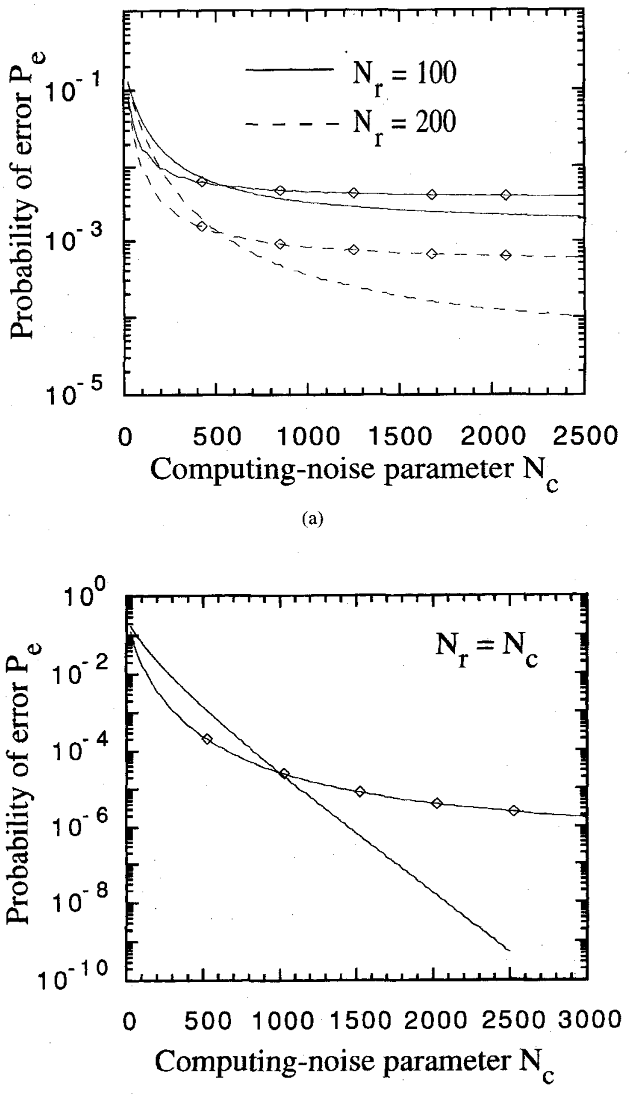
B. Selecting the Optical Threshold to Optimize the Performance

It is evident from our definition of the conditional probability of error that the choice of the threshold level may have an effect on the average probability of error. It is therefore important to investigate the possibility of optimizing the performance by a judicious selection of the threshold level. To motivate this point, consider the network in Example 1 and plot $P_e(\lambda, \mu)$ as a function of the optical threshold level ξ_1 over an admissible range $0.5 < \xi_1 < 1$ (see Fig. 3). The admissible range of a threshold level is the values of threshold that do not change the ideal characteristics of the network (see the hypothesis of the theorem). For fixed-noise parameters, there exists an optimum threshold level ξ_{optimum} at which $P_e(\lambda, \mu)$ is minimized. As the noise parameters increase, the optimal threshold decreases in value. This is expected since as the noise parameters increase, the decay of the tail of the probability density function of the doubly stochastic shot noise becomes faster. Furthermore, as the noise parameters increase, P_e becomes more sensitive to change in the optimum threshold value.

This procedure can be repeated for any single-layer network without difficulty if the nonlinearities are identical. On the other hand, if the thresholds ξ_i for any layer are allowed to be different, then the procedure becomes more cumbersome due to the fact that the optimization is performed over many threshold levels. It is therefore not generally mathematically tractable to determine the optimum threshold levels since the probability of error cannot be expressed in a closed form. It is possible, however, to use the asymptotic results to determine analytically the threshold levels that maximize the exponential decay rate r.

IV. PERFORMANCE OF MULTILAYER NETWORKS

Consider an M-layer network. For $k = 1, \cdots, M$, let $\mathbf{X}(k)$ and $\mathbf{X}(k+1)$ be vectors in $\{0,1\}^{d_k}$ and $\{0,1\}^{d_{k+1}}$, respectively, denoting the input and output to the kth layer.



(a)

(b)

Fig. 2. Average probability of error P_e for the single-layer networks in Example 1 as a function of the computing-noise parameter N_c: (a) The weight-recording-noise parameter N_r is fixed at 100 and 200 and (b) The weight-recording-noise parameter N_r is set to be equal to N_c. The curves labeled with diamonds correspond the network with 100 inputs, the remaining curves correspond to the two-input network.

The positive integer d_k, for each k, denotes the number of components $(\mathbf{X}(k)_i : i = 1, \cdots, d_k))$ of $\mathbf{X}(k)$. Let $S_k, (k = 1, \cdots, M+1)$ be a list of all the 2^{d_k} elements of $\{0,1\}^{d_k}$, and for each $i = 1, \cdots, 2^{d_k}$, let $I^i(k) = [I_1^i(k), \cdots, I_{d_k}^i(k)]'$ denote the ith item of the list S_k. The analysis of Section II enables us to compute the probabilities of the form

$$P_k(j \mid i) \triangleq \mathrm{P}\{\mathbf{X}(k+1) = I^j(k+1) \mid \mathbf{X}(k) = I^i(k)\}$$

where $i = 1, \cdots, 2^{d_k}$ and $j = 1, \cdots, 2^{d_{k+1}}$. Observe that

$$P_k(j \mid i) = \prod_{s=1}^{d_{k+1}} \mathrm{P}\{\mathbf{X}(k+1)_s = I_s^j(k+1) \mid \mathbf{X}(k) = I^i(k)\}$$

since the components $\mathbf{X}(k+1)_s$, $s = 1, \cdots, d_{k+1}$, of the vector $\mathbf{X}(k+1)$ are conditionally independent. Naturally, this setup gives rise for each $k = 1, \cdots, M$, to a $2^{d_k} \times 2^{d_{k+1}}$ transition probability matrix $\mathbf{T}(k)$ with entries

$$(\mathbf{T}(k))_{ij} = P_k(j \mid i).$$

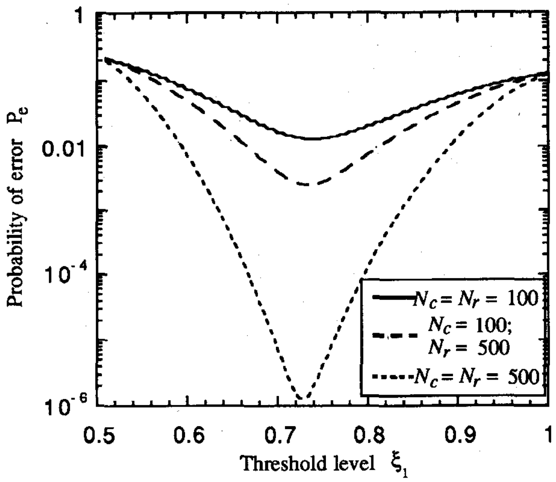


Fig. 3. The dependence of the average probability of error P_e on the selection of the threshold level ξ_1 for the two-input single-layer network of Example 1. Three sets of the computing-noise parameter N_c and the weight-recording-noise parameter N_r are considered: (100, 100), (100, 500), and (500, 500).

To determine the transition probability matrix of the M-layer network, we invoke the fact that for each $k = 2, \cdots, M+1$, the dependence of $\mathbf{X}(k)$ on $\mathbf{X}(1) \cdots \mathbf{X}(k-1)$ is through $\mathbf{X}(k-1)$ alone. In other words, the sequence $\{\mathbf{X}(k)\}_{k=1}^{M+1}$ is a Markov chain. In particular, $\mathrm{P}\{\mathbf{X}(M+1) = I^j(M+1) \mid \mathbf{X}(1) = \mathbf{x}_1, \cdots, \mathbf{X}(M) = I^i(M)\} = \mathrm{P}\{\mathbf{X}(M+1) = I^j(M+1) \mid \mathbf{X}(M) = I^i(M)\}$. The entries of the $2^{d_1} \times 2^{d_{M+1}}$ total probability transition matrix $\mathbf{T}_{\text{total}}$ are thus given by

$$\begin{aligned}(\mathbf{T}_{\text{total}})_{ij} &\triangleq \mathrm{P}\{\mathbf{X}(M+1) = I^j(M+1) \mid \mathbf{X}(1) = I^i(1)\} \\ &= (\mathbf{T}(1) \times \mathbf{T}(2) \times \cdots \times \mathbf{T}(M))_{ij}. \qquad (28)\end{aligned}$$

The average probability of error P_e is the probability of incorrect mapping averaged over all possible input patterns. If a network is aimed to implement a classifier, for instance, then P_e represents the average probability of any deviation from the desired classifier. For any input vector $\mathbf{x} \in S_1$, let $C(\mathbf{x}) \in S_{M+1}$ denote the desired output vector. Conditioned on a particular input pattern $\mathbf{X}(1) = \mathbf{x}$, the conditional probability of error $P_e(\mathbf{x})$ of the network is

$$\begin{aligned}P_e(\mathbf{x}) &\triangleq \mathrm{P}\{\mathbf{X}(M+1) \neq C(\mathbf{X}) \mid \mathbf{X}(1) = \mathbf{x}\} \\ &= 1 - \mathrm{P}\{\mathbf{X}(M+1) = C(\mathbf{x}) \mid \mathbf{X}(1) = \mathbf{x}\} \\ &= 1 - (\mathbf{T}_{\text{total}})_{i_1(\mathbf{x}), i_{M+1}(C(\mathbf{x}))}\end{aligned}$$

where $i_1(\mathbf{x})$ and $i_{M+1}(C(\mathbf{x}))$ are the indexes of $\mathbf{x}$ and $C(\mathbf{x})$ in S_1 and S_{M+1}, respectively. Hence

$$\begin{aligned}P_e &= \mathrm{E}[P_e(\mathbf{X}(1))] \\ &= \sum_{\mathbf{x} \in S_1} P_e(\mathbf{x}) \mathrm{P}\{\mathbf{X}(1) = \mathbf{x}\}.\end{aligned}$$

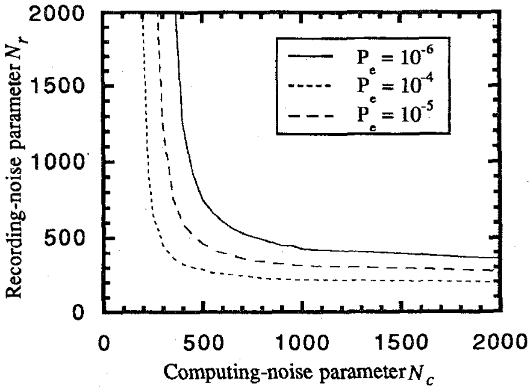


Fig. 4. Trade-off between the computing-noise parameter N_c and the weight-recording-noise parameter N_r for the three-layer classifier of Example 2. The average probability of error P_e is fixed at 10^{-4}, 10^{-5}, and 10^{-6}. For a given P_e, there is a critical value for the computing-noise parameter N_c below which the desired P_e is not attainable. A similar critical value for the weight-recording-noise parameter N_r is required.

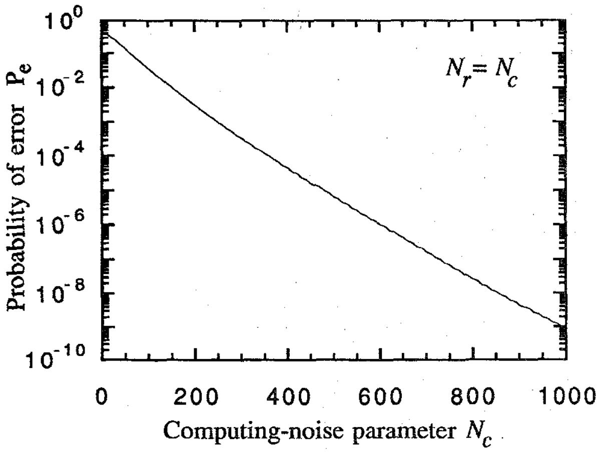


Fig. 5. Average probability of error P_e for the three-layer classifier in Example 2 as a function of the computing-noise parameter N_c. The parameter N_r is equal to N_c.

Example 2: Consider a three-layer classifier that maps the inputs "101," "011," and "110" to "1" and maps every other input to "0." All the temporal and spatial filters are chosen according to (25) and (26). Fig. 4 shows that to achieve a certain average probability of error P_e, the computing-noise parameter N_c and the weight-recording-noise parameter N_r must exceed certain levels. These levels are determined by the asymptotic behavior of P_e which can be computed by using the expression (19) and (18) for each layer. For example, if we require P_e to be 10^{-4}, then our calculations show that N_c and N_r should be at least 181. This demonstrates the trade off between the weight-recording-noise parameter N_r and the computing-noise parameter N_c, while fixing P_e. Clearly, for a lower P_e, we expect the curve to move up. Similarly to Example 1, as both N_r and N_c increase, P_e decreases to zero exponentially fast, as shown in Fig. 5. In this limiting case, the optical implementation of the network behaves identically to its deterministic counterpart.

V. PERFORMANCE OF RECURRENT NETWORKS

The performance analysis of the recurrent optical network follows directly from the analysis of the multilayer networks. The output at the kth iterate of the recurrent network can be thought of as the output of a k-layer neural network with all the layers having identical weight matrices and the same number of inputs and outputs. Let $\mathbf{T}$ denote the $m \times m$ one-step transition probability matrix of the network, then the $2^m \times 2^m$ matrix of k-step transition probabilities can be determined from (28) as $\mathbf{T}^k$. The conditional probability of error is $P_{e,k}(\mathbf{x}) = 1 - (\mathbf{T}^k)_{i(\mathbf{x}),i(C(\mathbf{x}))}$, where $i(\mathbf{x})$ and $i(C(\mathbf{x}))$ are the indexes of the initial state $\mathbf{x}$ and the desired output $C(\mathbf{x})$, respectively. The average probability of error $P_{e,k}$ for the kth iterate is then obtained by averaging over all $\mathbf{x}$ with respect to the initial distribution f.

A. Limit of Large Number of Iterations

Let the network's initial state $\mathbf{X}$ be set to $\mathbf{x}(1)$, and suppose that the network is designed so that a state $C(\mathbf{x})$ is achieved as the number of iterates tends to ∞. We wish to determine the probability that the state $C(\mathbf{x})$ is not attained, as the number of iterates tends to ∞.

Since this network is modeled by a finite state Markov chain with transition probability matrix $\mathbf{T}$, and since the entries of $\mathbf{T}$ are nonzero, as the number of steps (or iterates) increases, the probability that the final output of the network converges to a particular state $I_i \in \{0,1\}^m$ is independent of the initial state $\mathbf{X}(1)$, and these probabilities are the stationary distribution of the Markov chain. Let $\Pi = [\pi_1, \pi_2, \cdots, \pi_{2^m}]$ denote the stationary distribution of the Markov chain [9], i.e., Π is the unique nonnegative solution to the eigenvector equation $\Pi\mathbf{T} = \Pi$, with $\sum_i \pi_i = 1$. Furthermore, $\pi_i = \lim_{n\to\infty} P\{\mathbf{X}(n) = I_i\}$ regardless of the value of the initial state $\mathbf{X}(1)$. The probability of error $P_e(\mathbf{x})$ in sending the state $\mathbf{x}$ to the state $C(\mathbf{x})$ in an infinite number of iterations is simply

$$P_e(\mathbf{x}) = 1 - \pi_{i(C(\mathbf{x}))}$$

where $i(C(\mathbf{x}))$ denotes the index of the state $C(\mathbf{x})$ in the list S consisting of the elements of $\{0,1\}^m$. The average asymptotic probability of error P_e can be computed by averaging over $\mathbf{x}$ with respect to the initial probability mass function f, i.e., $P_e = \sum_{\mathbf{x}}(1 - \pi_{i(C(\mathbf{x}))})f(\mathbf{x})$.

Example 3: To illustrate the effect of the number of iterations on the performance of optical recurrent networks, consider the identity recurrent network whose weight matrix $\mathbf{W}$ is a 3×3 identity matrix. The threshold levels are set to 0.5. The temporal and spatial shot noise filters are assumed to be those given by (25) and (26). Fig. 6 demonstrates the dependence of the average probability of error $P_{e,k}$ on the number of iterations k along with asymptotic value P_e, as $k \to \infty$, for various values of the computing-noise parameter N_c and the weight-recording-noise parameter N_r. To achieve a certain accuracy level in a certain number of iterations, N_c and N_r must be chosen sufficiently large. For fixed values of these parameters, $P_{e,k}$ increases initially with k and then levels off to its asymptotic value. For this example, the asymptotic

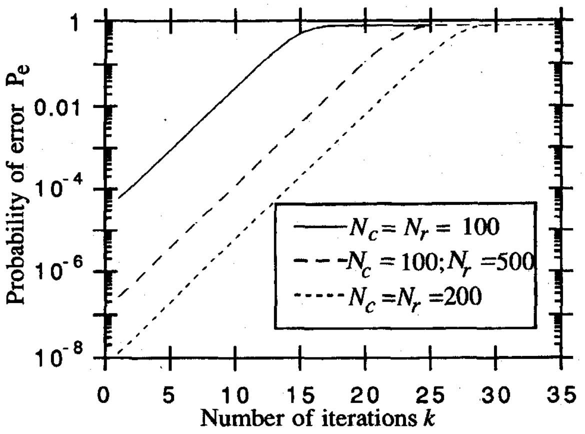


Fig. 6. The dependence of the average probability of error P_e on the number of iterations k for the 3 × 3 identity recurrent network of Example 3. Three sets of the computing-noise parameter N_c and the weight-recording-noise parameter N_r are considered: (100, 100), (100, 500), and (200, 200).

P_e is 0.875, and it is independent of the values of N_c and N_r. This is due to the one-to-one nature of this particular network.

VI. CONCLUSION

We have considered the performance analysis of neural networks for which weights and signals are modeled by shot-noise processes. Fluctuations in signals are referred to as computing noise and uncertainty in weights is referred to as weight-recording noise. This model is applicable to optical neural networks and biological systems in which signals have an inherent particle nature.

The dependence of the average probability of error in the network output has been determined in terms of key parameters of the computing noise and the weight-recording noise. The key parameter governing the statistics of the computing noise is the average number of quanta per clock cycle per signal. This parameter is referred to as the computing-noise parameter. The key parameter for the weight-recording noise is the average number of quanta per weight, and it is called the weight-recording-noise parameter. In an optical neural network, the computing-noise parameter is proportional to optical energy per processing time per beam; the weight-recording-noise parameter is proportional to the optical energy per pixel of spatial resolution.

We have shown analytically that for a fixed weight-recording-noise parameter, the probability of error decreases with the increase in the computing-noise parameter, and levels off to a value limited by the weight-recording-noise parameter. Similar behavior is obtained when the computing-noise parameter is fixed and the weight-recording-noise parameter is varied. For a given level of precision, there is therefore a trade off between weight-recording noise and computing noise. As the recording-noise parameter and the weight-computing-noise parameter are simultaneously increased, the average probability of error decays to zero exponentially fast as a function of the dominant parameter. The exponential decay rate was analytically determined. We found that the average probability of error can be minimized by an optimal selection of the nonlinearity thresholds. Furthermore, the sensitivity to this optimum threshold increases as the computing-noise parameter and the recording-noise parameter increase, i.e., the threshold robustness is lowered as a result of the reduction in computing and recording noise.

As for recurrent networks, we have captured the Markovian structure of the accumulation of noise from one iteration to the next. As the number of iterations increases, the average probability of error increases initially and then saturates at an asymptotic level. This level was characterized in terms of the stationary distribution of a Markov chain.

VII. APPENDIX: PROOF OF THE THEOREM

Without loss of generality, assume that $\theta_i > W_i\mathbf{x}$, $i = 1, \cdots, m$. In this case

$$P_e^i(\mathbf{x}, n) = \mathrm{P}\{Y_i \in (\xi_i, \infty) \,|\, \mathbf{X} = \mathbf{x}\}.$$

Let $\hat{Y} \overset{\triangle}{=} q_n Y_i$ where q_n is defined in Section III-A. Define

$$\tilde{Q}_n(s) \overset{\triangle}{=} q_n^{-1} \log \mathrm{E}[e^{s\hat{Y}} \,|\, \mathbf{X} = \mathbf{x}], \qquad s \in \mathrm{I\!R}.$$

Suppose that $\lim_{n\to\infty} \tilde{Q}_n(s) \overset{\triangle}{=} \tilde{Q}(s)$ exist and that it is differentiable. Put

$$I(u) = \sup_s \big(su - \tilde{Q}(s)\big), \qquad u \in \mathrm{I\!R}.$$

By Ellis's theorem [4]

$$\overline{\lim_{n\to\infty}}\, q_n^{-1} \log \mathrm{P}\left\{\frac{\hat{Y}_n}{q_n} \in [\xi_i, \infty) \,|\, \mathbf{X} = \mathbf{x}\right\} \le - \inf_{u\in[\xi,\infty)} I(u) \tag{29}$$

and

$$\underline{\lim_{n\to\infty}}\, q_n^{-1} \log \mathrm{P}\left\{\frac{\hat{Y}_n}{q_n} \in (\xi_i, \infty) \,|\, \mathbf{X} = \mathbf{x}\right\} \ge - \inf_{u\in(\xi,\infty)} I(u). \tag{30}$$

If I is continuous and increasing, then it follows that

$$\lim_{n\to\infty} q_n^{-1} \log \mathrm{P}\{Y \in (\xi_i, \infty) \,|\, \mathbf{X} = \mathbf{x}\} = -I(\xi_i).$$

Indeed, if condition 1) holds, then

$$\tilde{Q}(s) = k^{-1} \sum_{j=1}^{l} w_{ij} \beta(k x_j \alpha(s))$$

and calculus show that $I(u)$ is given by

$$I(u) = \rho_{1,i} u - k^{-1} \sum_{j=1}^{l} w_{ij} \beta(k x_j \alpha(\rho_{1,i}))$$

which is continuous and increasing, and Part 1) follows.

Parts 2) and 3) are proved similarly. Straightforward calculation shows that for Part 2)

$$\tilde{Q}(s) = \alpha(s)\gamma' W_i \mathbf{x}$$

and

$$I(u) = \rho_{2,i}u - \gamma'\alpha(\rho_{2,i})W_i\mathbf{x}.$$

For Part 3)

$$\tilde{Q}(s) = \sum_{j=1}^{l} w_{ij}\beta(\gamma s x_j)$$

and

$$I(u) = \rho_{3,i}\xi_i - \sum_{j=1}^{l} w_{ij}\beta(x_j\gamma\rho_{3,i}).$$

In either case, I is continuous and increasing, and Parts 2) and 3) of the theorem follow.

REFERENCES

[1] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation.* New York: Wiley 1990.

[2] M. Elbaum and M. Syrkin, "Perceptrons for photon-limited image classification," in *Proc. Soc. Photo-Optical Instrumentation Eng.*, vol. 1709, pp. 208–217, 1992.

[3] T. N. Cornsweet, *Visual Perception.* New York: Academic, 1970.

[4] R. S. Ellis, "Large deviations for a general class of random vectors," *Ann. Probability*, vol. 12, no. 1, pp. 1–12, 1984.

[5] N. H. Farhat, D. Psaltis, and E. Paek, "Optical implementation of the Hopfield model," *Appl. Opt.*, vol. 24, no. 10, 1985.

[6] A. F. Gmitro, P. E. Keller, and G. R. Gindi, "Statistical performance of outer-product associative memory models," *Appl. Opt.*, vol. 28, no. 10, pp. 1940–1948, 1989.

[7] J. A. Gubner and M. M. Hayat, "A method to recover counting distributions from their characteristic functions," *IEEE Signal Proc. Lett.*, to appear June 1996.

[8] T. L. Jong, "Noise effects and fault tolerance in Hopfield-type neural networks," Ph.D. dissertation, Texas Tech. Univ., Lubbock, TX, 1990.

[9] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes.* New York: Academic, 1975.

[10] R. P. Lippman, L. Kukolich, and E. Singer, "LNKnet: Neural networks, machine-learning, and statistical software for pattern classification," *Lincoln Lab. J.*, vol. 6, no. 2, 1993.

[11] B. E. A. Saleh, "Quantum fluctuations and adaptive spatial interactions in the visual system," *IEEE Trans. Syst., Man., Cybern.*, vol. SMC-8, pp. 875–879, 1978.

[12] B. E. A. Saleh and M. C. Teich, "Multiplication and refractoriness in the cat's retinal-ganglion-cell discharge at low light levels," *Biol. Cybern.*, vol. 52, pp. 101–107, 1985.

[13] B. E. A. Saleh, "Noise in nonlinear optical computing," in *Proc. Annu. Mtg. IEEE Lasers Electro-Opt. Soc.*, San José, CA, Nov. 1991.

[14] ———, "Quantum noise in optical processing," in *Real-Time Optical Information Processing*, B. Javidi and J. Horner, Eds. New York: Academic, 1994.

[15] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space.* New York: Springer-Verlag, 1991.

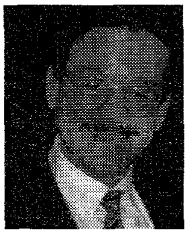
[16] M. Stevenson, R. Winter, and B. Widrow, "Sensitivity of feedforward neural networks to weight errors," *IEEE Trans. Neural Networks*, vol. 1, no. 1, pp. 71–80, 1990.

[17] L. Zhang, "Impact of noise on an optoelectronic neural network," *SPIE*, vol. 1709, pp. 629–639, 1992.

Majeed M. Hayat (S'89–M'92) was born in Kuwait in 1963. He received the B.S. degree in 1985 (summa cum laude) in electrical engineering from the University of the Pacific, Stockton, CA. He received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Wisconsin-Madison in 1988 and 1992, respectively.

He is currently a Postdoctoral Research Associate and a Lecturer at the Department of Electrical and Computer Engineering at the University of Wisconsin-Madison. His research interests include statistical signal processing, quantum image estimation, performance of optical communication systems and photonic networks, and point processes applied to spatial pattern analysis.

Dr. Hayat is a Member of the Optical Society of America, Eta Kappa Nu, Phi Kappa Phi, and Tau Beta Sigma.



Bahaa E. A. Saleh (M'73–SM'86–F'91) received the B.S. degree from Cairo University, Egypt, in 1966 and the Ph.D. degree from the Johns Hopkins University, Baltimore, MD, in 1971, both in electrical engineering.

From 1971 to 1974 he was Assistant Professor at the University of Santa Catarina, Brazil, and from 1974 to 1976 he was Research Associate at the Max Planck Institute in Göttenberg, Germany. In 1977–94 he was on the faculty of the Department of Electrical and Computer Engineering at the University of Wisconsin-Madison, and he served as Department Chairman in 1990-94. He is currently Professor and Chairman of the Electrical, Computer, and Systems Engineering Department at Boston University, MA. He held visiting appointments at the University of California-Berkeley and Columbia University, NY. His research interests include optics and photonics including statistical and quantum optics, optical signal processing, optical communication, nonlinear optics, photodetectors, digital image processing, and vision. He is the author of two books, *Photoelectron Statistics* (New York: Springer-Verlag, 1978) and *Fundamentals of Photonics* (New York: Wiley, 1991), the coeditor of *Transformations in Optical Signal Processing* (Bellingham, WA: Soc. Photo-Optical Instrumentation Eng., 1983), and the Author of chapters in seven books. He is the author or coauthor of more than 150 papers in technical journals.

Dr. Saleh is the Editor-in-Chief of the *Journal of the Optical Society of America A*, a Member of the board of editors of the *Journal of the European Optical Society B: Quantum Optics*, and Coeditor of the *Adam Hilger Optics and Optoelectronics Series*, (U.K.: Institute of Physics). He is a Fellow of the Optical Society of America and a member of Phi Beta Kappa and Sigma Xi. In 1981 he received the University of Wisconsin Romnes Award, and in 1984 he became Fellow of the Guggenheim Foundation.

John A. Gubner (S'83–M'85) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1983, 1985, and 1988, respectively.

In 1988 he joined the faculty of the University of Wisconsin-Madison, where he is now an Associate Professor in the Department of Electrical and Computer Engineering. From 1986 to 1988 he was a Graduate Fellow with the Systems Research Center at the University of Maryland. His research interests include information theory, shot-noise random processes, non-Poisson point processes, Markov chain Monte Carlo, distributed estimation, and wavelets.

In 1985 Dr. Gubner received an IEEE Frank A. Cowan Scholarship for graduate study in communications.