Foundations of A Computational Theory of Catecholamine Effects

Harry Printz
David Servan-Schreiber

May 21, 1990

CMU-CS-90-105 3

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This report presents the mathematical foundation of a theory of catecholamine effects upon human signal detection abilities. We argue that the performance-enhancing effects of catecholamines are a consequence of improved rejection of internal noise within the brain.

To support this claim, we develop a neural network model of signal detection. In this model, the release of a catecholamine is treated as a change in the gain of a neuron's activation function. We prove three theorems about this model. The first asserts that in the case of a network that contains only one unit, changing its gain cannot improve the network's signal detection performance. The second shows that if the network contains enough units connected in parallel, and if their inputs satisfy certain conditions, then uniformly increasing the gain of all units does improve performance. The third says that in a network where the output of one unit is the input to another, under suitable assumptions about the presence of noise along this pathway, increasing the gain improves performance. We discuss the significance of these theorems, and the magnitude of the effects that they predict.

This research was supported by the Defense Advanced Research Projects Agency (DOD) and monitored by the Space and Naval Warfare Systems Command under Contract N00039-87-C-0251, ARPA Order No. 5993 .

Keywords: neural networks, signal detection, catecholamines, gain, constant optimal performance, ensemble effect, chain effect

1 Introduction

The catecholamines are a group of neuroactive substances that are believed to modulate information processing in the brain, as opposed to transmitting particular pieces of information. Much is known about their effect upon the responsivity of individual neurons, and upon gross aspects of behavior like alertness and attention. However, there is as yet no comprehensive theory that explains the latter effects in terms of the former.

In this paper we advance the mathematical foundations for such a theory. We believe that the effects of catecholamines upon human performance at signal detection tasks are a consequence of improved rejection of noise within the brain. They may also arise partly from an enhancement in the ability of the perceptual apparatus to extract a signal from noisy sensory data, though this effect appears to be small.

Our argument proceeds along the following lines. We construct a model of signal detection in biological systems, which captures certain known features of neurons, and ignores others. Then we prove several properties of this model, and use these formal results to explain the gross behavioral impact of catecholamines.

We have described this as a computational theory, which no doubt gives some indication of the nature of our model. Specifically, we treat the neuron as a device that computes a real-valued function of its net input, called an *activation function*, and we treat the influence of the catecholamines as a change in this function. We will not concern ourselves with the biochemical machinery that implements the activation function, nor with the precise way in which catecholamines affect this mechanism. Our work is an investigation of the performance of an assembly of these devices, commonly called a *neural network*, upon a particular computational task. This task is a formalization of a real signal detection task performed by human subjects.

The formal results we prove about our model consist of three theorems. Each one concerns the effect of changes in the activation function upon the signal detection performance of a neural network.

The first of these, the Constant Optimal Performance Theorem, asserts that if a network contains only one unit, changing its gain cannot improve the network's signal detection performance. This result is significant because it undermines explanations of catecholamine effects that are formulated exclusively in terms of the signal-to-noise ratio.

The second, the Ensemble Performance Theorem, shows that if the network contains a number of units connected in parallel, and if their inputs satisfy certain conditions, then uniformly increasing the gain of all units does improve the network's performance. This amounts to enhancing the network's ability to extract a signal from noisy inputs, a phenomenon we call the *ensemble effect*. However, our numerical studies suggest that the magnitude of this effect is small.

The third, the Chain Performance Theorem, says that in a network where the output of one unit supplies the input to another, under suitable assumptions about the presence of noise along this connection, increasing the gain improves performance. As we will see, this means that the network is doing a better job at rejecting internal

noise. We call this the *chain effect.*

The apparent contradiction between the first two results is resolved below. Its resolution depends in an essential way upon summing the outputs of several units. Thus one corollary of this work is the demonstration of a truly emergent property of neural networks.

Note that these results are all statements about our model. They do not touch on how well the model accords with real biological systems, or to what extent the behavior of the model explains the effects of catecholamines upon human subjects. These issues are addressed in a companion paper [12].

The plan of the current paper is as follows. In Section 2 we review the neurophysiological phenomena that motivated this study. In Section 3, we discuss the neural network models to which the theorems apply, describe the signal detection task the networks perform, and introduce the terminology and notation we will use throughout the paper. In Section 4 we state the theorems and explain what they say, and provide some intuition about why they are true. In Sections 5, 6 and 7 we prove the Constant Optimal Performance Theorem, the Ensemble Performance Theorem, and the Chain Performance Theorem respectively. In Section 8 we give a critical review of the different explanations we have provided for catecholamine effects, and suggest directions for future research. In Section 9 we summarize our results and contributions.

The reasoning presented here draws upon results in probability theory and real analysis. We have tried to keep the descriptive material at an introductory level. Sections 2, 3, 4, 8 and 9 require only elementary probability theory, at the level of [14, Sections 9.1–9.8], and some previous exposure to signal detection theory, such as [6, Chapter 1]. To read and understand the proofs requires a working knowledge of real analysis at the level of [10] and [17], and familiarity with the results of [16, Chapters 2 and 3].

2 Neurophysiological Motivation

In this section we outline the neurophysiological phenomena that motivated our work. First we discuss the catecholamines, and evidence that these substances modulate neural responsivity. Then we describe the phenomenon that we propose to explain, which is the enhancement, with catecholamine release, of human signal detection performance. This outline is not intended as a critical evaluation of competing hypotheses about the role of the catecholamines, but merely as a sketch of one widely held view, which motivated our investigations.

The catecholamines are a group of neuroactive substances, consisting of dopamine and its metabolic products, norepinephrine and epinephrine [3]. Dopamine and norepinephrine were originally thought to function as *inhibitory neurotransmitters;* that is, their release at a synapse was thought to reduce the firing rate of the postsynaptic neuron [3, p 267]. However, more recent studies suggest that these substances may act as *neuromodulators.* By "neuromodulator" we mean a substance that does

not itself alter the firing rate of a neuron, but which changes the cell's response to putative neurotransmitters.

There is now a body of evidence suggesting that dopamine and norepinephrine function in this way. For example, the presence of dopamine at a synapse in the striatum, at concentrations too low to change the basal firing rate of the postsynaptic neuron, increases both the excitatory effect of the neurotransmitter glutamine (GLU), and the inhibitory effect of the neurotransmitter γ-aminobutyric acid (GABA) [1]. That is, upon release of GABA in the presence of dopamine, there is a greater reduction of the neuron's firing rate from its basal level than if the same amount had been released in the absence of dopamine. Likewise for the firing rate increase induced by the release of GLU. For reasons that will become apparent later, we will refer to this modulatory effect as *raising the neuron's gain*. Similar results apply to norepinephrine [18]. Moreover, once a catecholamine has been released, its influence on the target cell may last several seconds or even minutes, whereas the effects of GABA and GLU last only a few milliseconds.

In addition to these pharmacological results, there is anatomical and physiological evidence that the catecholamines function as gross modulators of information processing within the brain, rather than as messengers of particular neural signals.

The anatomy of systems of catecholamine-containing neurons allows them to influence neuronal activity throughout much of the neocortex. In primate brains, nearly all of the cell bodies of noradrenaline-containing neurons are found in the locus ceruleus, a small, well-circumscribed nucleus. While the bodies and preterminal axons of these neurons contain relatively low concentrations of catecholamines, the varicosties at the ends of the axons contain very high concentrations [3, pp 259–261]. These axons project to many cortical structures, including the primary somatosensory cortex, the primary motor cortex, the primary visual cortex, and area 7 of the parietal lobe. In addition, fibers from the locus ceruleus innervate almost all subcortical structures, including the thalamus. Dopamine-containing neurons originate in the substantia nigra and the ventro tegmental area, and project to subcortical motor structures, the primary motor cortex, and the associational neocortex of the prefrontal, temporal and parietal lobes.

In contrast to thalamus-to-cortex and cortex-to-cortex fibers, which project to sharply delineated areas of the cortex, catecholaminergic fibers branch in profusion. Moreover, whereas thalamic fibers project radially, leading more or less directly from the thalamus to their destination, catecholaminergic fibers follow along the surfaces of brain structures. A single axon from a catecholaminergic neuron may traverse several functionally distinct cortical regions. This anatomical organization is well suited to global modulation, but not to spatially localized transmission of signals.

The catecholaminergic systems also have physiological properties that set them apart from other types of neural systems. Their conduction velocity is slow, and their baseline firing rate is low and stable, resulting in a steady release of catecholamines. These neurons exhibit a constricted firing range, and are not able to sustain high levels of activity. Their apparent function is to maintain a constant level of norepinephrine and dopamine at their axon terminals. Finally, at least in the locus ceruleus, when this

body is stimulated, all its neurons fire uniformly, independent of the specific source of stimulation. It appears to weigh inputs from its two or possibly three afferents, and then widely distribute a uniform chemical message. In fact, the modulatory effects of norepinephrine that we described above have also been induced by the stimulation of the locus ceruleus [18].

Now we turn to the effect of catecholamine release upon human information processing. Our measure of this effect is the continuous performance test. In one version of this test, a subject watches a sequence of letters flashed for fixed duration one after another upon a screen, and is asked to press a button whenever any two successive letters are the same. Two identical consecutive letters are said to be an instance of the *target event*. Pressing the button when there was no target event is called a *false alarm*; failing to report a target event is called a *miss*.

We compare the performance of subjects before and after receiving a central nervous system stimulant, such as amphetamine, that directly releases catecholamines from synaptic terminals, and blocks their reuptake. Before medication, typical performance is 15% to 20% misses, and 0.5% to 1% false alarms. After medication, the fraction of misses drops to 6% to 12%, while the fraction of false alarms is unchanged [8]. We take this as evidence that one possible effect of the catecholamines upon the human information processing system is to improve the ability of subjects to extract a signal from a noisy background. These results have been closely matched by a computer simulation of neural behavior during this task [13].

Our aim is to explain the improvement in signal detection performance of systems of neurons in terms of the effect of the catecholamines upon individual neurons. Several researchers [5,18] have attempted to account for these observations in terms of the "improved signal-to-noise ratio" of the individual cell. Unfortunately, these accounts are formulated too imprecisely to have explanatory value. For example, they do not specify where noise enters the cell. As we shall see, this has a substantial influence upon whether or not the modulatory effect of the catecholamines can actually improve signal detection performance. Moreover, signal-to-noise ratio (SNR) is not a characteristic of the cell, but of the input incident upon it, or output emerging from it. Probably what these researchers had in mind was the ratio between the input and output SNRs. But it turns out that even increasing this ratio may not improve performance.

In this paper we attempt to meet these difficulties head-on. We formulate models of neural behavior, and of signal detection performance, that are mathematically precise, yet broad enough to encompass real biological systems. Reasoning formally about these models, we establish the existence of the ensemble effect and the chain effect. We provide numerical examples of both effects; these examples have led us to conclude that stimulant-induced enhancement of signal detection performance is primarily a consequence of the chain effect. We also point out some gaps in our explanations. These are areas where somewhat more information (about the function of real biological systems) and insight (into the mathematical behavior of our models of these systems) are required, before we can claim to have a complete understanding of these phenomena.

3 Network Models and Signal Detection

In this section we establish the framework for our theorems. We say what kinds of networks they apply to, describe the signal detection task we are modeling, and introduce some terminology and notation.

3.1 Network Models

Our work concerns three different kinds of neural networks, which we call *single-unit*, *multi-unit* and *chain*. These are illustrated in Figure 1 below. The circles are

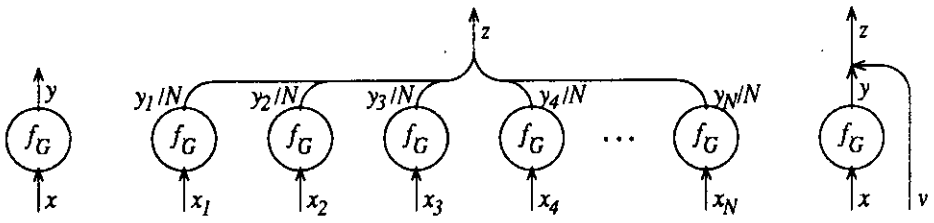

Figure 1: Single-Unit, Multi-Unit and Chain Networks

"units," which are intended to model the action of neurons. Each unit works in the same way: it takes its net input x, which is a real number, computes the value $f_G(x)$, which is guaranteed to lie in the range $(0, 1)$, and supplies this as its output y. We will sometimes refer to y as the unit's *activation;* it represents a neuron's firing rate, or equivalently, its probability of firing.

The only difference between single-unit and multi-unit networks is that there are N identical units in the multi-unit case. Here the N individual outputs $y_1, \ldots, y_N$ are each multiplied by $1/N$, and summed to yield z, the network output. Note that because of the weighting by $1/N$, the multi-unit output z is also guaranteed to lie in $(0, 1)$. The only difference between single-unit and chain networks is that an *output noise* term v is added to the unit's activation y to yield the final output z. This means that z is no longer guaranteed to lie in $(0, 1)$.

f_G is called the unit's *activation function*. It is a strictly increasing function from the reals to $(0, 1)$. We put no conditions on its continuity or differentiability. The G subscript is meant to indicate that the particular function f_G is drawn from a *family* of activation functions, $\{f_G\}$. Each value of the *gain* parameter G, where $G > 0$, determines a strictly increasing function from the reals to $(0, 1)$. For instance, the set of biased logistic functions, given by

$$f_G(x) = \frac{1}{1 + e^{-(Gx-1)}}$$

is such a family. Figure 2 below shows the graphs $y = f_G(x)$ for two members of this family.

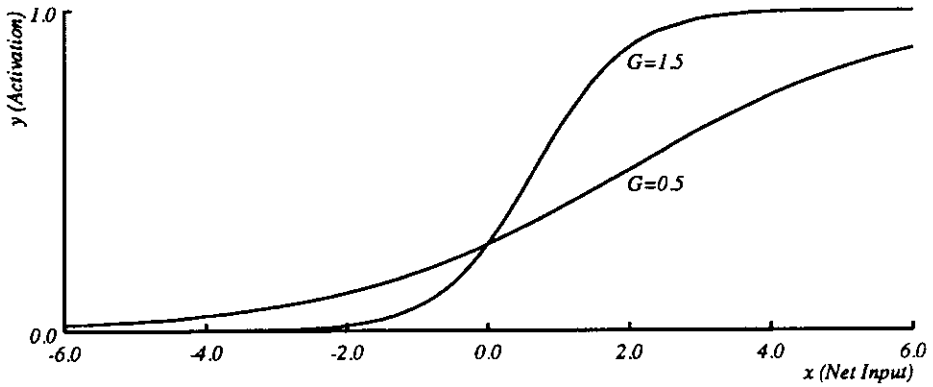

Figure 2: Two Members of a Typical Activation Family

There is one additional condition that we impose on the family $\{f_G\}$. We require that as $G \to \infty$, the function f_G converges pointwise almost everywhere to the unit step function u_0, where u_0 is defined by

$$u_0(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

and that this convergence is monotone increasing for $x \geq 0$, and monotone decreasing for $x < 0$. This means that as G increases, the value $f_G(x)$ gets steadily closer to 1 if $x \geq 0$, and steadily closer to 0 if $x < 0$, except possibly on a negligible set of points. For a rigorous discussion of convergence almost everywhere, consult [17] or [11]. We will say that a set of functions $\{f_G\}$ that satisfies this, plus the earlier conditions on each member function, is an *activation family*. As an example, for any fixed value of the bias B, the biased logistics $f_G(x) = 1/(1 + e^{-(Gx+B)})$ are an activation family.[1]

3.2 The Signal Detection Task

Now we describe the signal detection task that we want these networks to perform. We begin by reviewing the notion of signal detection in general, then specialize this to the single-unit and multi-unit networks.

When we speak of "signal detection," we mean the following. Suppose there are two possible situations, or "states of the world," that we wish a receiver to distinguish. We call these two states "signal present" and "signal absent"—a bit of a misnomer, since in both cases there is some input to the receiver. It is the job of the receiver to process this input, determine the state of the world, and announce "detect" when the signal is present, or "ignore" when it is absent.

[1]The logistics are used throughout purely as a familiar example. The theorems we develop apply to *any* activation family.

Now let the receiver be a single-unit network, operating at a fixed gain G. Suppose for a moment that its input can take on only two values: x_S for "signal present," and x_A for "signal absent," with $x_A < x_S$. We wish to discriminate between these two cases by observing the output y. Since f_G is strictly increasing, we can do this by selecting a threshold θ that satisfies $f_G(x_A) < \theta < f_G(x_S)$. Then to determine the state of the input, we compare the output y with θ. If $y > \theta$, we announce "detect," if $y < \theta$, "ignore." We call the first case a *hit*, the second, a *correct ignore*.

This task is easy because we have assumed that the input is noise-free—it is always either x_S or x_A, right on the mark. But this is never the case in any real-world situation, since the input will often be corrupted by noise.

We can capture the effect of noise by modeling the input with probabilities. The input to the unit is now a random variable (hereafter "rv") that is described by its probability density function (pdf). In the presence of signal, it is the rv X_S, described by pdf ρ_{X_S}. In the absence of signal, it is the rv X_A, with pdf ρ_{X_A}. We require that these pdfs are Lebesgue integrable [17], but we impose no other conditions on them. For those who are unfamiliar with Lebesgue integration, there is a brief discussion of the subject at the end of this section.

Since the input is now described by a pdf, so too is the output. X_S and X_A determine new rvs, $Y_S = f_G(X_S)$ and $Y_A = f_G(X_A)$, which represent the output of the unit in the presence and absence of signal respectively. We write ρ_{Y_S} and ρ_{Y_A} for their pdfs, which are determined by the action of f_G upon ρ_{X_S} and ρ_{X_A}. The situation is summarized in Figure 3, which shows a typical activation function f_G, input pdfs ρ_{X_S} and ρ_{X_A}, and the resulting ρ_{Y_S} and ρ_{Y_A}.

If it happens that ρ_{Y_S} and ρ_{Y_A} do not overlap, then we can continue to distinguish perfectly between presence and absence of signal. But if they do overlap, we are bound to make some mistakes. Referring to Figure 3, suppose we select $\theta = 1/2$ as threshold. It is apparent that both

$$\int_0^{\frac{1}{2}} \rho_{Y_S}(\xi)\, d\xi \quad \text{and} \quad \int_{\frac{1}{2}}^1 \rho_{Y_A}(\xi)\, d\xi$$

are non-zero. These are respectively the probability that the output falls below threshold with signal present, and the probability that it exceeds threshold with signal absent. Thus there is a non-zero chance that we will ignore the signal when it is in fact present—called a *miss*—or conversely "detect" it when it is actually absent—a *false alarm*.

Unfortunately, when the pdfs overlap, the problem of misses and false alarms will arise no matter what value we choose for θ. The best we can do is determine a threshold θ^* that is optimal in the following sense. Let us assign a payoff to each of the four possible situations in our signal detection task. These are listed in the following payoff matrix

	detect	ignore
signal present	D	$-M$
signal absent	$-F$	I

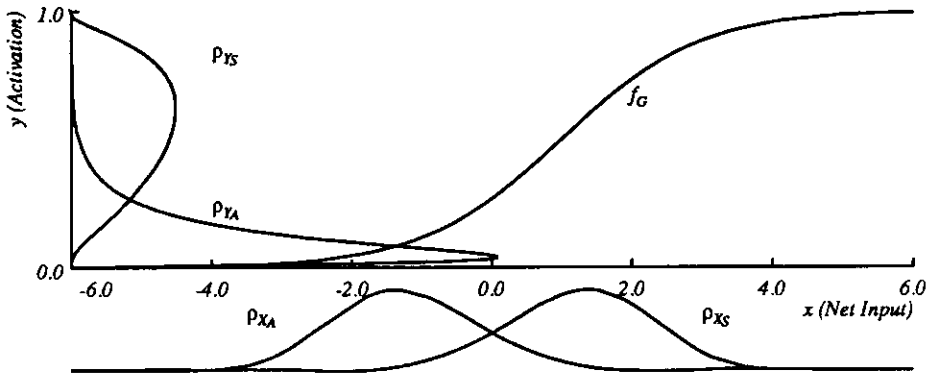

Figure 3: Input and Output Probability Density Functions. The curves at the bottom are the pdfs of the net input in the signal absent (left) and signal present (right) cases. Thus the probability that the input lies in $[x_0, x_1]$ in each case is the area under the corresponding curve between these limits. The curves along the y-axis are the transformed pdfs for each case; they are functions of the activation y, and represent the distribution of outputs. Thus the probability that the output lies in $[y_0, y_1]$ is the area of the region bounded below by a line at y_0, on the right by the appropriate pdf, above by a line at y_1, and on the left by the y-axis. Note: the X and Y pdfs are drawn to different scales.

where D, M, F and I are non-negative. Thus we are rewarded an amount D for correctly detecting signal present, and I for ignoring signal absent, but penalized F for a false alarm, and M for a miss.

If we write P_S and P_A for the prior probabilities of signal presence or absence, then it is not hard to show [6] that the *expected payoff* at the threshold θ, written $E(\theta)$, is given by

$$E(\theta) = \lambda + \alpha \int_\theta^1 \rho_{Y_S}(\xi)\, d\xi - \beta \int_\theta^1 \rho_{Y_A}(\xi)\, d\xi,$$

where

$$\begin{aligned} \lambda &= I \cdot P_A - M \cdot P_S, \\ \alpha &= (D + M) \cdot P_S \quad \text{and} \\ \beta &= (F + I) \cdot P_A. \end{aligned}$$

Note that α and β are both non-negative. We will refer to the value $E(\theta)$ as the network's *performance* at threshold θ, and to the function E as the *performance function*.

By solving the equation $dE/d\theta = 0$ we can determine a value $\theta^\star$ that maximizes E. (Note that E is a continuous function on the compact domain $[0, 1]$, and therefore attains a maximum.) Let us write $E^\star$ for $E(\theta^\star)$, the expected payoff at optimal threshold. We say that $\theta^\star$ is the *optimal threshold*, and $E^\star$ is the network's *optimal performance*, on the signal detection task that is determined by the given pdfs, prior probabilities and payoffs. This means that if we use $\theta^\star$ to discriminate between presence and absence of signal, though we will still make some mistakes, we will maximize the long-run expected payoff.

We can now say precisely what we mean by a signal detection task for a single-unit network. Such a task $\mathcal{T}$ is specified by

$P_S, P_A,$	the prior probabilities of the signal present and signal absent states of the world,
$D, M, F, I,$	non-negative payoffs, as in the matrix above, and
$\rho_{X_S}, \rho_{X_A},$	Lebesgue integrable pdfs of the input in the signal present and signal absent states.

If either α or β equals 0, then we will say $\mathcal{T}$ is a *trivial* signal detection task. For if $\alpha = 0$, then clearly $\theta^\star = 1$, and if $\beta = 0$, then $\theta^\star = 0$, independent of ρ_{Y_S} and ρ_{Y_A}. If α and β are both non-zero, we will say $\mathcal{T}$ is *non-trivial*.

The case when the receiver is a multi-unit network is a straightforward extension of these ideas. We consider a network of a fixed number of units N, connected as in Figure 1, each operating at the same fixed gain G. But now we have N inputs $x_1, \ldots, x_N$. In the presence of signal, these inputs are described by N independently distributed rvs $X_{S1}, \ldots, X_{SN}$, each with pdf ρ_{X_S}, and in the absence of signal by $X_{A1}, \ldots, X_{AN}$, each with pdf ρ_{X_A}.

The input rvs are each transformed by f_G exactly as in the single-unit case. Thus in the presence of signal the outputs are described by N independent rvs $Y_{S1}, \ldots, Y_{SN}$, each with pdf ρ_{Y_S}. These are weighted and summed to yield the network output z, which is an rv

$$Z_S = \frac{Y_{S1} + \cdots + Y_{SN}}{N}$$

with pdf ρ_{Z_S}. Since the summands are independent rvs, ρ_{Z_S} is the convolution of N copies of $\rho_{(Y_S/N)}$, the pdf of Y_S/N. Likewise in the absence of signal we have the N rvs $Y_{A1}, \ldots, Y_{AN}$, each with pdf ρ_{Y_A}, which yield the output rv

$$Z_A = \frac{Y_{A1} + \cdots + Y_{AN}}{N}$$

with pdf ρ_{Z_A}.

As in the single-unit case, the multi-unit receiver must distinguish between the signal present and absent states by comparing the network output z with a threshold θ. But if ρ_{X_S} and ρ_{X_A} overlap, so too will ρ_{Z_S} and ρ_{Z_A}, and we once again have the problem of misses and false alarms. As before, the performance of the network at

threshold θ is given by

$$E(\theta) = \lambda + \alpha \int_{\theta}^{1} \rho_{Z_S}(\xi)\, d\xi - \beta \int_{\theta}^{1} \rho_{Z_A}(\xi)\, d\xi,$$

where α, β and λ are unchanged. And as before, we can find the optimal threshold θ^*, thereby determining $E^* = E(\theta^*)$, the optimal performance of the multi-unit network on the signal detection task.

The only difference between the signal detection task for multi-unit and single-unit receivers is that there are N inputs instead of one. Thus we will adopt the same definition for a multi-unit signal detection task $\mathcal{T}$, where it is understood that all the input rvs have the same pdf ρ_{X_S} or ρ_{X_A}, depending upon the state of the world.

An identical development may be carried through for a chain. We model the output noise term ν by the random variable V, with pdf ρ_V. The network's output in the presence of noise is an rv $Z_S = Y_S + V$; in the absence of noise it is $Z_A = Y_A + V$. Hence the performance function is

$$E(\theta) = \lambda + \alpha \int_{\theta}^{\infty} \rho_{Z_S}(\xi)\, d\xi - \beta \int_{\theta}^{\infty} \rho_{Z_A}(\xi)\, d\xi.$$

Since we are treating the output noise as a property of the network, and not of the signal detection task, we can adopt the single-unit definition of $\mathcal{T}$ unchanged.

To close this section, since some readers may be unfamiliar with Lebesgue integration, we will now provide a brief discussion of what is entailed when we say a pdf is Lebesgue integrable. Suppose X is a real-valued random variable, and ρ_X is its pdf. Then the probability that X lies in a set D, written $\Pr(X \in D)$, is given by the integral

$$\int_{-\infty}^{\infty} \chi_D(\xi) \cdot \rho_X(\xi)\, d\xi,$$

where $\chi_D(\xi)$ is a function that equals 1 if $\xi \in D$, and 0 otherwise.

Imagine now that there were some single real value r such that the probability that X has precisely this value, and no other, is non-zero. If this were so, then the value of $\Pr(X \in D)$ would change as we adjoined or deleted the single point r from D. In such cases we say that X has an *atom* at r. It is as if there were an infinitely small, indivisible hunk of "probability," located right at r.

However, our intuition is that nothing takes place with infinite precision in real biological systems. Suppose the rv X represents the value of some physical parameter—say the membrane potential of a neuron. There is no one single real value such that the probability that X has precisely that value, and no other, is non-zero. So we want to exclude the situation just described. Thus we require that X has no atoms; saying that ρ_X is Lebesgue integrable implies this.

This same intuition underlies our requirement that the activation function f_G be strictly increasing. For suppose to the contrary that there were a number $y \in [0, 1]$, and a non-empty interval (x_1, x_2), such that for each x in (x_1, x_2), we have $f_G(x) = y$.

Consider the rv $Y = f_G(X)$. Then by definition,

$$\Pr(Y = y) = \Pr(f_G(X) = y).$$

But $f_G(X)$ attains the value y at least whenever X lies in (x_1, x_2), and possibly in other cases as well. So $\Pr(Y = y) \geq \Pr(X \in (x_1, x_2))$. Thus if the right-hand quantity is non-zero—which is entirely possible, whether or not X has atoms—then Y has an atom at y. It is for this reason that we require f_G to be strictly increasing for all G. Note that any increasing function can be approximated uniformly, with arbitrary precision, by a strictly increasing function, so this requirement does not impose any practical restrictions.

It is possible to develop a concept of the integral, known as the Stieltjes integral, for dealing with rvs that have atoms. In fact, we will make use of the Stieltjes integral in the proof of the Ensemble Performance Theorem. But even when we use this concept, we will require that X is atomless and f_G is strictly increasing, for the plausibility reasons just cited.

One more comment about integrals. So far we have explicitly exhibited the variable of integration, as in

$$\int_{-\infty}^{\infty} \rho_X(\xi)\, d\xi.$$

But this is really a useless piece of notation, unless the integrand is a multivariate function, or the variable of integration appears elsewhere in the integrand. For this reason, from now on we drop the "$(\xi)\, d\xi$," except in special cases.

3.3 Notation for Gain

In the preceding discussion we made no mention of the gain G, except to say that it was fixed. We did this to clarify the discussion of the performance function, the optimal threshold, and the performance at optimal threshold. However, in what follows we will be greatly concerned with the effect of changing the gain. In particular, we will need to exhibit the value of the gain in effect when certain quantities are computed. We now extend the notation of the previous section to do this.

We begin with the random variables X_S and X_A, with their pdfs ρ_{X_S} and ρ_{X_A}. These characterize the inputs to the network in the two possible states of the world. Since these quantities are inputs, by assumption they are not affected by variations in the gain.

Each unit passes its input through the activation function f_G to determine its output, and here the gain enters the picture. We write Y_{GS} for the random variable $f_G(X_S)$, the output of a unit *operating at the gain G* in the signal present case. Y_{GS} is described by its pdf, which we write as $\rho_{Y_{GS}}$. In general, the shape of $\rho_{Y_{GS}}$ changes with G—see Figure 4 for an example of this dependence. Naturally, the mean of $\rho_{Y_{GS}}$ will also depend upon G. We write

$$\mu(Y_{GS}) = \int_0^1 \xi \cdot \rho_{Y_{GS}}(\xi)\, d\xi$$

for this statistic. The symbols Y_{GA}, $\rho_{Y_{GA}}$ and $\mu(Y_{GA})$ denote similar quantities for the signal absent case. Likewise, we have rvs Z_{GS} and Z_{GA}, with pdfs $\rho_{Z_{GS}}$ and $\rho_{Z_{GA}}$, and means $\mu(Z_{GS})$ and $\mu(Z_{GA})$ for the output of the multi-unit network. Finally, we write $\sigma^2(V)$ for the variance of any random variable V.

4 What the Theorems Mean

We are now in a position to state and discuss the three main results of this paper: the Constant Optimal Performance Theorem, the Ensemble Performance Theorem and the Chain Performance Theorem. Their proofs are somewhat technical, and we defer them to later sections. Here we formulate the theorems, say what they mean, and supply some intuition about why they are true.

4.1 The Constant Optimal Performance Theorem

This theorem is an assertion about the dependence—or rather, the independence—of the optimal performance of a single-unit network upon the gain. Let's return to the signal detection task $\mathcal{T}$ for this network, taking care to exhibit just how the gain enters into the determination of $\theta^\star$ and $E^\star$.

Since the pdfs ρ_{X_S} and ρ_{X_A} of the input are part of the *definition* of the task facing the network, they are of course independent of the gain. However, the pdfs $\rho_{Y_{GS}}$ and $\rho_{Y_{GA}}$ of the output rvs Y_{GS} and Y_{GA} are decidedly *not* independent of the gain. This is evident in Figure 4, which shows f_G, $\rho_{Y_{GS}}$ and $\rho_{Y_{GA}}$ for three different values of G. In view of the dramatic dependence of the output pdfs upon the gain, it is natural to ask if changing G can improve the network's signal detection performance.

Let us formulate this question more precisely. We are given a particular signal detection task $\mathcal{T}$, specified by P_S and P_A; D, M, F and I; and ρ_{X_S} and ρ_{X_A}. For any fixed value of the gain G, we can determine the performance function

$$E_G(\theta) = \lambda + \alpha \int_\theta^1 \rho_{Y_{GS}} - \beta \int_\theta^1 \rho_{Y_{GA}},$$

where α, β and λ do not depend upon G. Let $\theta_G^\star$ be the threshold for which this function attains its maximum. Does there exist a different value of the gain, G', with some possibly different optimal threshold $\theta_{G'}^\star$, such that $E_{G'}^\star$, the optimal performance at G', is greater than $E_G^\star$, the optimal performance at G? More simply put, we want to know if we can find G' such that

$$E_{G'}^\star > E_G^\star.$$

The Constant Optimal Performance Theorem answers this question in the negative. We state the theorem now, and prove it below in Section 5.

> **Theorem (Constant Optimal Performance)** *Let $E_G^\star$ be the performance at optimal threshold of a single-unit network on the signal detection*

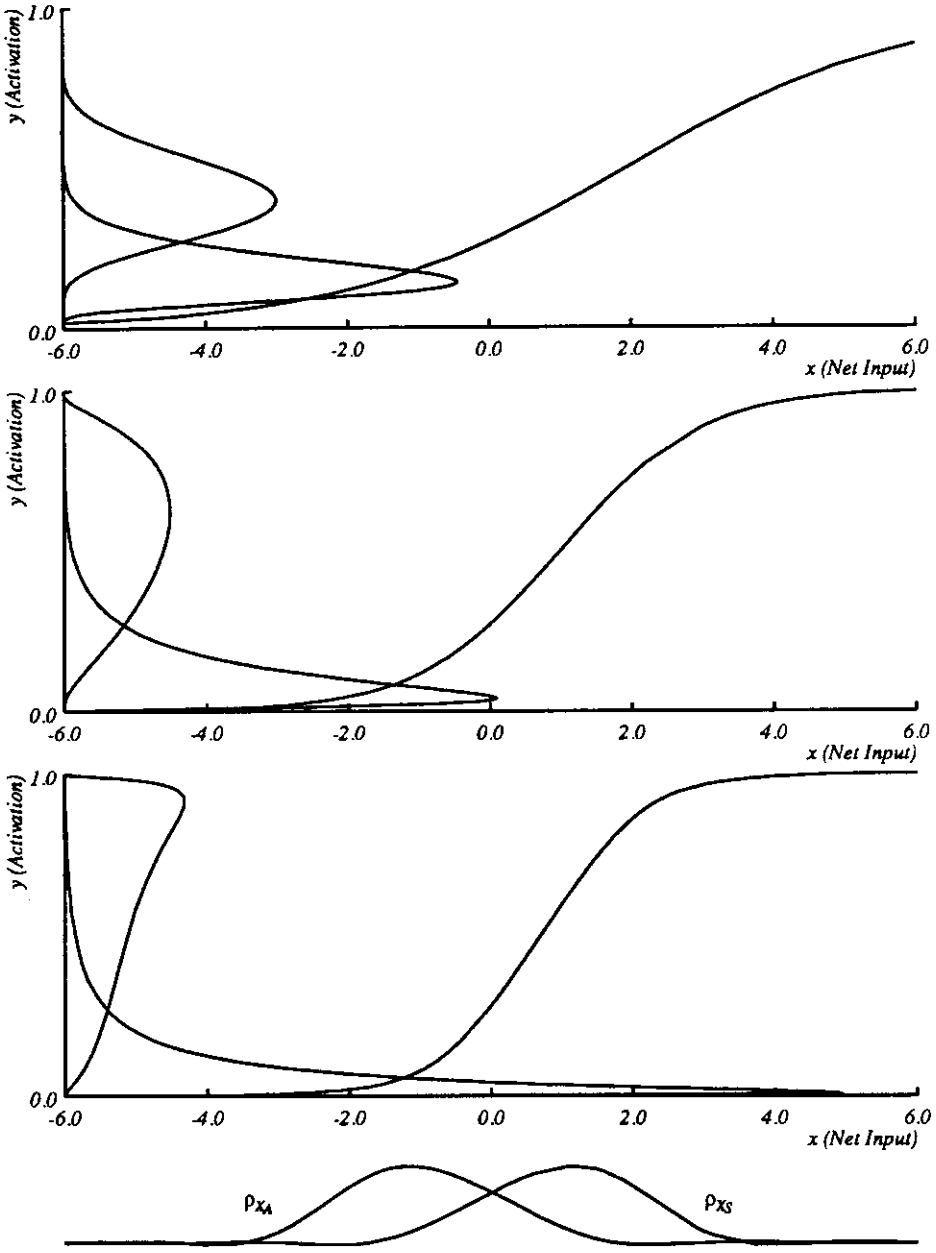

Figure 4: Dependence of Output Pdfs Upon Gain. These graphs use the same conventions and input pdfs as Figure 3. They depict the biased logistic f_G and the output pdfs for the gain values 0.5, 1.0 and 1.4 (top to bottom). The bias B was fixed at -1 throughout.

task T, and let $\{f_G\}$ be the unit's activation family. Then E^*_G is a constant, independent of the gain G.

It is important to understand just what this theorem means. It does not say that $E_G(\theta)$ is independent of the gain. In general, at any fixed threshold θ, the performance can go either up or down as we increase the gain. (Of course, if we are at θ^*_G, the performance can only go down.) Nor does it say that there is some fixed threshold θ^* that is optimal for all gains. The situation is summarized in Figure 5 below. This figure displays a hypothetical performance function for three different values of the gain, $G_1 < G_2 < G_3$. By considering the threshold $\theta = 1/2$, for example, we see that

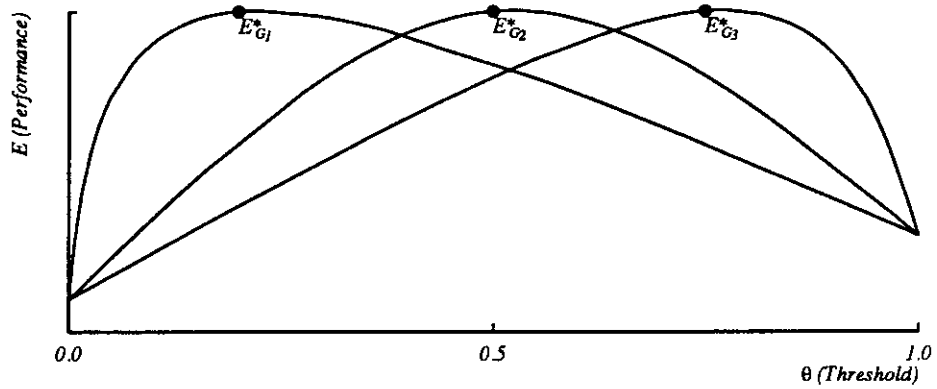

Figure 5: Performance as a Function of Threshold for Three Different Gains

the performance may rise and then fall with increasing gain. The key observation is that the maxima $E^*_{G_1}$, $E^*_{G_2}$ and $E^*_{G_3}$, attained respectively at thresholds $\theta^*_{G_1}$, $\theta^*_{G_2}$ and $\theta^*_{G_3}$, are all exactly the same.

We want to underscore the generality of this theorem. It imposes absolutely no conditions on the original signal present and signal absent pdfs, ρ_{X_S} and ρ_{X_A}, except that they are Lebesgue integrable. These of course must be fixed; we are not claiming that the unit's optimal performance is the same for wholly different signal detection tasks. But for *any* two fixed ρ_{X_S} and ρ_{X_A} whatsoever, the theorem will hold. We have used Gaussian pdfs for the input rvs in our figures, but that is only because this is a familiar distribution, and a likely one for naturally occurring noise.

Readers familiar with signal detection theory may wonder why we have not couched our discussion in terms of the parameter d', which is often used as a measure of the distinguishability of signal-present and signal-absent cases [6, p. 60]. This is an important point. The parameter d' is the difference of the means of the output rvs Y_{GS} and Y_{GA}, divided by their standard deviation, *under the assumption that these rvs are both Gaussian, with a common standard deviation.* But as we have taken some pains to show in our figures, the output pdfs have a complicated internal structure,

which is not fully captured by their mean and standard deviation alone. Increasing the gain may in fact drive apart the means $\mu(Y_{GS})$ and $\mu(Y_{GA})$. But this does not imply that their pdfs $\rho_{Y_{GS}}$ and $\rho_{Y_{GA}}$ are being rigidly translated away from one another. Such an intuition is false and misleading.

This point is significant because the average power output of a neuron can be shown to be proportional to the mean of its firing rate. Now the signal-to-noise ratio (SNR) through a channel is defined as the quotient of the average power through the channel in the presence of signal, and the average power in absence of signal. Hence even though increasing the gain may raise the SNR at the output of a single-unit network, its performance at optimal threshold will remain constant. Thus any theory that accounts for the performance effects of the catecholamines exclusively in terms of the SNR is inadequate. But there is much that is right about such theories, and we will investigate them further in Section 8.

We can say a bit about why the theorem is true, without going into the details of the proof. Suppose we try to determine the optimal threshold by differentiating $E_G(\theta)$ with respect to θ, and setting the derivative equal to zero. We have

$$\begin{aligned} \frac{dE_G}{d\theta} &= \alpha \cdot \frac{d}{d\theta} \int_\theta^1 \rho_{Y_{GS}} - \beta \cdot \frac{d}{d\theta} \int_\theta^1 \rho_{Y_{GA}} \\ &= -\alpha \cdot \rho_{Y_{GS}}(\theta) + \beta \cdot \rho_{Y_{GA}}(\theta). \end{aligned}$$

Thus $\theta_G^\star$ satisfies

$$\alpha \cdot \rho_{Y_{GS}}(\theta_G^\star) = \beta \cdot \rho_{Y_{GA}}(\theta_G^\star). \qquad (\dagger)$$

Now

$$\Pr(Y_{GS} \geq \theta) = \Pr(f_G(X_S) \geq \theta) = \Pr(X_S \geq f_G^{-1}(\theta)).$$

Assume for a moment that f_G is differentiable, with a differentiable inverse f_G^{-1}. (The actual proof imposes no such restrictions.) From this it is possible to show that

$$\rho_{Y_{GS}}(\theta) = \rho_{X_S}(f_G^{-1}(\theta)) \cdot f_G^{-1\prime}(\theta),$$

and likewise for $\rho_{Y_{GA}}$. Hence (†) may be rewritten

$$\alpha \cdot \rho_{X_S}(f_G^{-1}(\theta_G^\star)) = \beta \cdot \rho_{X_A}(f_G^{-1}(\theta_G^\star)).$$

This means that $\theta_G^\star$ satisfies (†) if and only if $x^\star = f_G^{-1}(\theta_G^\star)$ satisfies

$$\alpha \cdot \rho_{X_S}(x^\star) = \beta \cdot \rho_{X_A}(x^\star). \qquad (\dagger\dagger)$$

But now observe that equation (††) is *completely independent of the gain*. Its solution $x^\star$ is determined by the signal detection task $\mathcal{T}$, independent of any activation function f_G. The only effect of a gain change is to move $\theta_G^\star = f_G(x^\star)$ around in $(0, 1)$.

Now the optimal performance at gain G is given by

$$E_G^\star = \lambda + \alpha \cdot \Pr(Y_{GS} \geq \theta_G^\star) - \beta \cdot \Pr(Y_{GA} \geq \theta_G^\star).$$

But for the probability of a hit at optimal threshold, we have

$$\Pr(Y_{GS} \geq \theta_G^\star)\Pr(f_G(X_S) \geq \theta_G^\star) = \Pr(X_S \geq f_G^{-1}(\theta_G^\star)) = \Pr(X_S \geq x^\star),$$

and likewise for the probability of a false alarm at optimal threshold, $\Pr(Y_{GA} \geq \theta_G^\star)$. That is, though adjusting G alters $\theta_G^\star$, this value moves in just such a way as to leave unchanged the chances of both a hit and a false alarm at optimal threshold. Therefore for a fixed signal detection task $\mathcal{T}$,

$$E_G^\star = \lambda + \alpha \cdot \Pr(X_S \geq x^\star) - \beta \cdot \Pr(X_A \geq x^\star)$$

is a constant.

These insights have another important consequence. Suppose we alter the payoffs and prior probabilities of $\mathcal{T}$, but leave ρ_{X_S} and ρ_{X_A} unchanged. We are varying $\alpha = (D + M) \cdot P_S$ and $\beta = (F + I) \cdot P_A$. This means that for fixed ρ_{X_S} and ρ_{X_A}, we can move $x^\star$ anywhere we like on the real axis, and hence $\theta_G^\star$ in $(0, 1)$, just by manipulating the prior probabilities P_S and P_A (or if we like, the entries in the payoff matrix). This is illustrated in Figure 6 below.

In particular, note that the analogous equation in the multi-unit case,

$$\alpha \cdot \rho_{Z_{GS}}(\theta) = \beta \cdot \rho_{Z_{GA}}(\theta), \qquad (\ddagger)$$

obtained by differentiating the expression for the multi-unit performance with respect to θ, is satisfied by the optimizing $\theta_G^\star$. This means that $\theta_G^\star$ can be moved where we please in $(0, 1)$ by altering the prior probabilities or the payoffs. This will have important consequences for the applicability of the Ensemble Performance Theorem.

The alert reader may be wondering why, if equation (‡) determines $\theta_G^\star$ for a multi-unit receiver, the Constant Optimal Performance Theorem cannot be extended to cover this case as well. The answer is that in the single-unit case, the simple relationships

$$\Pr(Y_{GS} \geq \theta) = \Pr(X_S \geq f_G^{-1}(\theta)) \quad \text{and} \quad \Pr(Y_{GA} \geq \theta) = \Pr(X_A \geq f_G^{-1}(\theta))$$

permit us to reduce the equation for $E^\star$ to one in which the gain makes no appearance. But in the multi-unit case, no such reduction is possible. We know that $\rho_{Z_{GS}} = \rho_{(Y_{GS}/N)}^{*N}$, where the ^{*N} denotes the N-fold convolution of the pdf with itself. But when we try to express

$$\Pr(Z_{GS} \geq \theta) = \int_\theta^1 \rho_{(Y_{GS}/N)}^{*N}$$

in terms of ρ_{X_S}, we find that f_G is inextricably woven into $\rho_{(Y_{GS}/N)}^{*N}$, and likewise for $\rho_{(Y_{GA}/N)}^{*N}$. This prevents us from finding a gain-independent expression for the performance. Similar difficulties arise when we attempt to formulate equation (‡) in a gain-independent way. These problems make it impossible to carry the argument through.

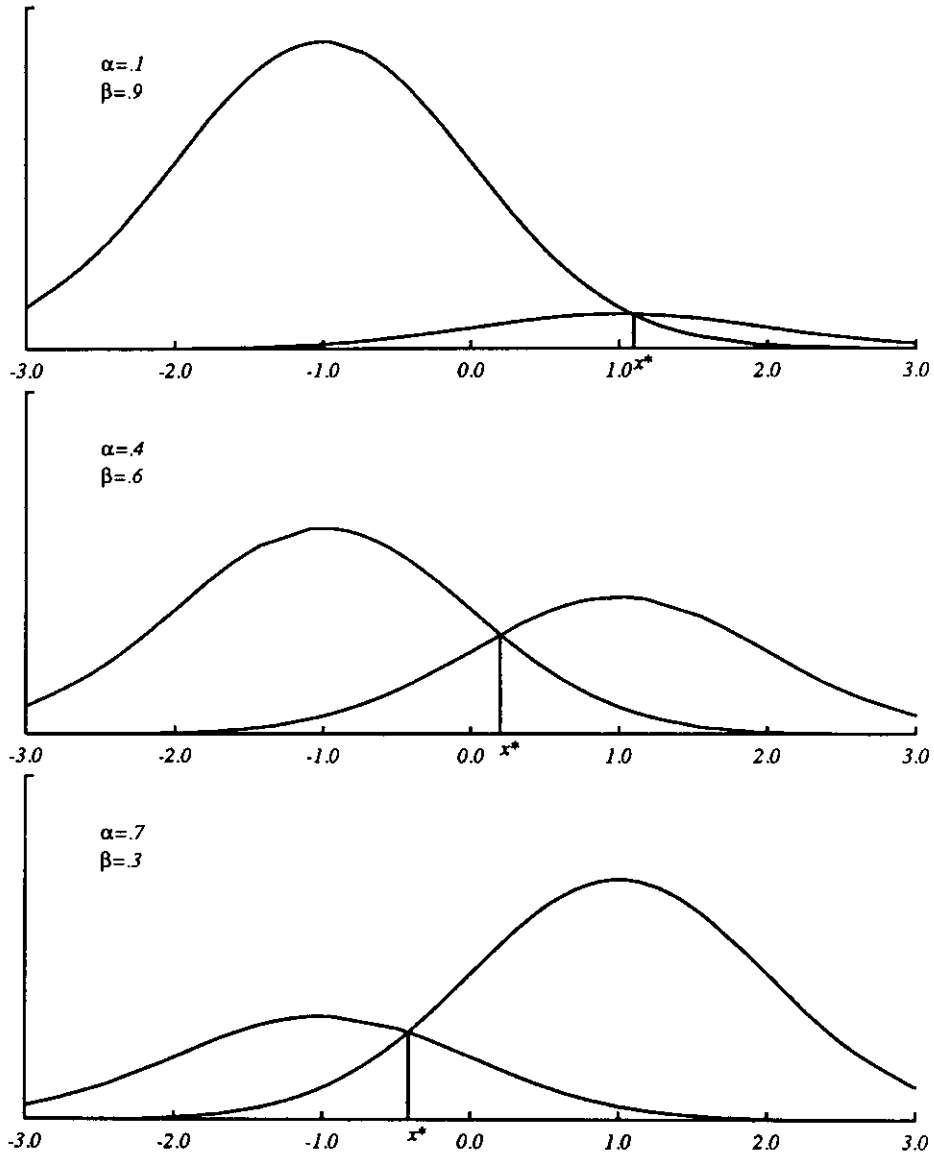

Figure 6: Variation of x^* With α and β. This figure demonstrates how changing the prior probabilities and the payoffs moves x^* around on the real line, and hence moves $\theta_G^* = f_G(x^*)$, which is not shown here, around in $[0, 1]$. In each of the graphs above, the lefthand curve is $\beta \cdot \rho_{X_A}$, and the righthand curve is $\alpha \cdot \rho_{X_S}$. Their intersection determines the solution x^* of the equation $\alpha \cdot \rho_{X_S}(x) = \beta \cdot \rho_{X_A}(x)$. In these examples, we have taken payoffs so that $\alpha = P_S$ and $\beta = P_A$, and hence $\alpha + \beta = 1$, but this of course is not a requirement.

4.2 The Ensemble Performance Theorem

The Ensemble Performance Theorem asserts that under suitable conditions, the optimal performance of a multi-unit receiver can be improved by increasing the gain. No doubt this sounds like the exact opposite of the Constant Optimal Performance Theorem. We resolve this apparent logical contradiction below.

The aim of this section is to formulate this result, and provide some intuition about why it holds, without giving all the details of the proof. Our approach will be to retrace the reasoning that led us to the theorem, introducing the appropriate concepts as they arise.

This discussion parallels that of the preceding section. But it is almost twice as long, and for this reason we divide it into four subsections. First we provide the foundations and intuitions that underlie this result. Then we supply a worked example of the fundamental concept. Next we state the theorem. Finally we discuss its meaning and give a numerical example.

4.2.1 Foundations and Intuitions

We begin with two basic conditions that plainly must be fulfilled if we are to have any hope at all of proving a theorem that says what we desire. These concern the number of units N, and the signal detection task $\mathcal{T}$.

First, in view of our previous result, the new theorem cannot possibly hold for a single-unit receiver. Thus the theorem must include some sort of lower bound on N. Second, the signal detection task $\mathcal{T}$ must be non-trivial, since otherwise no change in G, or for that matter ρ_{X_S} or ρ_{X_A}, will alter the performance.

Next, we take note of two simple conditions on the relation between X_S and X_A that must be part of the hypotheses of the theorem, either explicitly or implicitly. The first is that they must not be identical. For if they were, we could not possibly discriminate between the signal present and signal absent states, no matter what the value of the gain. The second is that X_S and X_A must not be separable. By this we mean there must be no x^* for which both $X_A < x^*$ always and $X_S > x^*$ always. For if this were so, at any gain G we could choose $\theta_G^* = f_G(x^*)$ and discriminate without error between the two states, and therefore could not improve the performance by increasing the gain.

Anecdotally, these conditions say that the situation must neither be so apallingly bad, nor so astoundingly good, that we cannot hope to improve it. As we shall see, both of them turn out to be consequences of the hypotheses of the theorem.

Next we will proceed to investigate, in very general terms, the effect of increasing the number of units. But we want to emphasize that our result concerns the improved performance, with increasing gain, of a network with a *fixed* number of units N. We investigate the effect of increasing N only because it helps understand some of the concepts that arise in the proof.

This said, we proceed with our study. Let's consider the signal absent case. We assume that the gain is fixed at some value G, and write Y_A for Y_{GA}, likewise Z_A for

Z_{GA}, and so on. Recall that Z_A is defined as

$$Z_A = \frac{\sum_{i=1}^{N} Y_A}{N}.$$

Then we have at once $\mu(Z_A) = \mu(Y_A)$, so increasing N has no effect upon the mean.

As N increases, Z_A tends to a Gaussian random variable, centered on $\mu(Y_A)$, with a peak that becomes sharper and more narrow with increasing N. This is the meaning of the central limit theorem, and it is the basis of the statistical result that taking more independent samples reduces the chance of error. It is straightforward to show that $\sigma^2(Z_A) = \sigma^2(Y_A)/N$, which gives us a measure of how rapidly this peak narrows with increasing N. Figure 7 is an example of how the pdf of the sum varies with the number of summands. In the figure, each summand is the rv $Y_A = f_G(X_A)$. The input rv X_A and gain G are fixed; only N is changing.

Note that for any fixed $\theta > \mu(Y_A)$, the tail probability $\Pr(Z_A \geq \theta)$ diminishes as N increases. This probability, which is the chance of a false alarm, equals the area beneath the curve ρ_{Z_A} and to the right of θ. As the peak narrows and sharpens, more and more probability mass is concentrated near the mean, so this area shrinks as N gets larger.

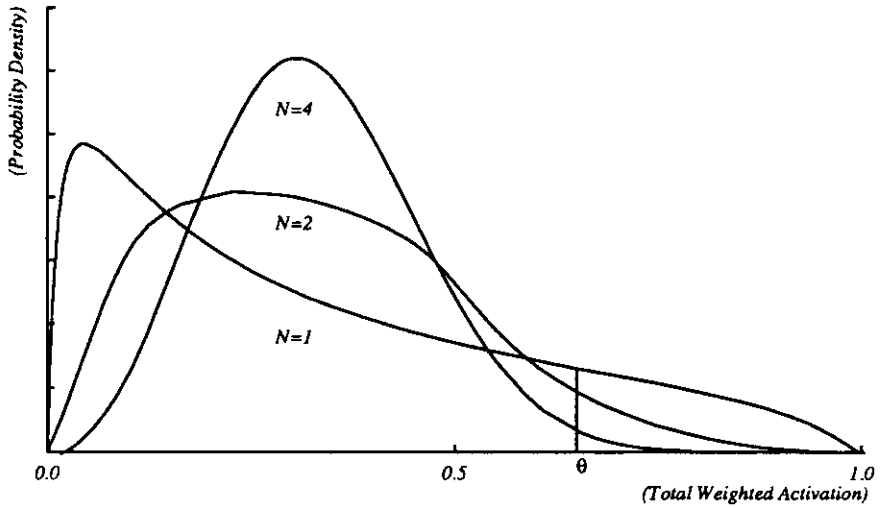

Figure 7: Pdfs for Three Values of N. The curves shown are the pdfs of the rv $Z_A = \sum_{i=1}^{N} Y_A/N$, where each rv $Y_A = f_G(X_A)$. The area to the right of θ and below the curve represents the tail probability, $\Pr(Z_A \geq \theta)$. Note how this probability shrinks with increasing N.

It can be shown [16] that providing $\theta > \mu(Y_A)$, as N increases $\Pr(Z_A \geq \theta)$ falls to 0 like ${\gamma_A}^N$. Here γ_A is a real number in $[0, 1]$ that depends upon θ and the shape

of ρ_{Y_A}; it measures the rate at which the tail probability decreases. We say that γ_A is the *rate number* for the rv Y_A. This property is expressed precisely by the limit equation

$$\gamma_A = \lim_{N \to \infty} \Pr(Z_A \geq \theta)^{1/N}, \qquad (\S)$$

which we will take as the definition of γ_A.

Since γ_A depends upon θ and Y_A, we should write $\gamma(\theta, Y_A)$ for the rate number. And since Y_A itself depends upon X_A and f_G, it would be still more proper to write $\gamma(\theta, X_A, f_G)$. But in the interest of simplifying notation we will generally drop these reminders and ask the reader to keep the dependence in mind.

Some readers may be wondering about the requirement that $\theta > \mu(Y_A)$. We have imposed it to ensure that $\Pr(Z_A \geq \theta)$ actually does fall as $N \to \infty$. For if $\theta \leq \mu(Y_A)$, then as N increases, $\Pr(Z_A \geq \theta)$ *rises*, to a limiting value of 1. We can ease this restriction by taking $\gamma_A = 1$ if $\theta \leq \mu(Y_A)$, and by saying not that the tail "falls to 0 like ${\gamma_A}^N$," but only that it "behaves like ${\gamma_A}^N$" as N increases. This is consistent with our definition of γ_A, and simply amounts to recognizing that (§) applies more widely than we originally expected.

There is another issue that we want to mention now, though it cannot be as easily resolved. In writing equation (§), we have assumed that the limit exists. But there is no obvious reason why this should be so. It is entirely conceivable that even though the tail falls as N increases, it does so in a way that does not conform asymptotically to any familiar function.

Remarkably, the limit in (§) exists under extremely broad conditions—so broad that for the purposes of this work, we need impose no additional restrictions upon the input pdfs, or the activation functions that transform them. In Section 6 we will demonstrate this, and discuss how to compute the rate number from a pdf, and how it relates to more familiar concepts. For the moment it is enough to note that the limit exists, and that the rate number for Z_A depends upon the gain. This completes our discussion of the effect of increasing N.

Now let us restore the G subscript to our notation, and investigate the effect of increasing the gain. We consider two values of the gain, G and G', with $G < G'$. If we set $Y_{GA} = f_G(X_A)$ and $Y_{G'A} = f_{G'}(X_A)$ then the random variables

$$Z_{GA} = \frac{\sum_{i=1}^{N} Y_{GA}}{N} \quad \text{and} \quad Z_{G'A} = \frac{\sum_{i=1}^{N} Y_{G'A}}{N}$$

represent the network output at these two gain values. We write γ_{GA} and $\gamma_{G'A}$ for the respective rate numbers of these rvs, where

$$\gamma_{GA} = \lim_{N \to \infty} \Pr(Z_{GA} \geq \theta)^{1/N} \quad \text{and} \quad \gamma_{G'A} = \lim_{N \to \infty} \Pr(Z_{G'A} \geq \theta)^{1/N}.$$

Suppose now that we could show that $\gamma_{G'A} < \gamma_{GA}$; then clearly ${\gamma_{G'A}}^N < {\gamma_{GA}}^N$ for all N. Now roughly speaking, for large N the chance of a false alarm $\Pr(Z_{GA} \geq \theta)$ lies close to ${\gamma_{GA}}^N$, and likewise for $\Pr(Z_{G'A} \geq \theta)$ and ${\gamma_{G'A}}^N$. (This is not quite what

the definition (§) says; getting round this difference requires a mathematical trick that we defer to the proof.)

In particular, we can find an integer N, possibly very large but nonetheless fixed and finite, such that $\Pr(Z_{G'A} \geq \theta) < \Pr(Z_{GA} \geq \theta)$. *This means that for a network of N units, we can cause the probability of a false alarm to drop by increasing the gain from G to G'.*

We have couched this discussion in terms of two fixed values of the gain, where the initial G and a higher G' are both known. In general though the situation takes the following less constrained form. Given the initial gain G, can we find some higher gain that gives improved performance? This leads us to formulate the discussion in terms of the limit of the rate number as G increases without bound. We write this as $\gamma_{\infty A}$, defined by $\gamma_{\infty A} = \lim_{G \to \infty} \gamma_{GA}$. Then by an extension of the reasoning of the preceding paragraph, we have that for a suitably large network, providing $\gamma_{\infty A} < \gamma_{GA}$, we can reduce the probability of a false alarm by increasing the gain from G to a sufficiently large value.

We give this property a special name. Fix a threshold θ. We say that ρ_{X_A} is a *gain improvable density* (gid) if, for some fixed value of N, we can reduce the tail probability from $\Pr(Z_{GA} \geq \theta)$ to a smaller value $\Pr(Z_{G'A} \geq \theta)$ by increasing the gain from G to G'. If we want to underscore a particular threshold for which ρ_{X_A} has this property, we will say that it is a gid *for the threshold* θ. In view of the preceding paragraph, $\gamma_{\infty A} < \gamma_{GA}$ is a sufficient condition to make ρ_{X_A} a gid. This nomenclature extends to the signal present case as well, with the change that we consider the opposite tail, $\Pr(Z_{GS} \leq \theta)$.

4.2.2 A Gain Improvable Density

So far we have explained the concept of a gid, but we have not exhibited one. We now give an example of a gid, and also provide the promised intuition about why increasing the gain yields an improvement in the tail probability of the sum, but not of an individual unit. The discussion that follows lies at the heart of understanding our whole result, and the reader is urged to pay specially close attention.

Let X be a binomial random variable that attains the value $x_1 < 0$ with probability $1-p$, and the value $x_2 > 0$ with probability p. Then $Y_G = f_G(X)$ is likewise a binomial, attaining $y_1 = f_G(x_1)$ with probability $1-p$, and $y_2 = f_G(x_2)$ with probability p. We proceed to examine the probability that the sum $Z_G = \sum_{i=1}^{N} Y_G/N$ exceeds a threshold θ, where $y_1 < \theta < y_2$. Note that $\Pr(Y_G \geq \theta) = p$.

Z_G is a discrete random variable, which can attain only the $N+1$ values

$$\frac{Ny_1}{N}, \frac{(N-1)y_1 + y_2}{N}, \ldots, \frac{(N-i)y_1 + iy_2}{N}, \ldots, \frac{Ny_2}{N}$$

with probabilities

$$\binom{N}{0}(1-p)^N p^0, \binom{N}{1}(1-p)^{N-1}p^1, \ldots, \binom{N}{i}(1-p)^{N-i}p^i, \ldots, \binom{N}{N}(1-p)^0 p^N$$

respectively. Thus we may find an explicit expression for $\Pr(Z_G \geq \theta)$ as follows. We determine the least integer J_G such that

$$\frac{(N - J_G)y_1 + J_G y_2}{N} \geq \theta;$$

this may be obtained as

$$J_G = \left\lceil N \cdot \frac{\theta - y_1}{y_2 - y_1} \right\rceil.$$

Then we have

$$\Pr(Z_G \geq \theta) = \sum_{i=J_G}^{N} \binom{N}{i} (1-p)^{N-i} p^i.$$

We will make use of this quantity shortly.

As the gain increases without bound, y_1 falls to 0, and y_2 rises to 1. So in the limit, which we denote by the ∞ subscript, we have a random variable $Y_\infty = u_0(X)$, which is just a binomial that attains 0 with probability $1 - p$, and attains 1 with probability p. As before, we define $Z_\infty = \sum_{i=1}^{N} Y_\infty / N$.

Let's compare Y_∞ with Y_G. By increasing the gain, we have not altered the probability that an individual unit exceeds the threshold; this remains fixed at p. But when Y_∞ does exceed θ, it does so by a greater amount than Y_G would have. Likewise, when Y_∞ lies below θ, as it does with probability $1 - p$, it falls short by more.

Now we make a key observation. Unlike the tail probability of an individual rv like Y_G or Y_∞, when we consider the sum Z_∞, the margin by which each summand Y_∞ / N exceeds or falls short of θ / N is important, since such excesses or shortfalls will be accumulated together, and thereby may determine whether or not the total exceeds the threshold. Let us attempt to reason about the effect of a gain increase upon Z_G. We expect a fraction of about p of the summands to increase their contribution, but we expect a fraction $1 - p$ to decrease. Hence providing the decrease sufficiently exceeds the increase, the tail probability of Z_∞ will likely lie below the tail probability of Z_G.

We now justify this intuition. Proceeding as before, we obtain

$$\Pr(Z_\infty \geq \theta) = \sum_{i=J_\infty}^{N} \binom{N}{i} (1-p)^{N-i} p^i,$$

where

$$J_\infty = \left\lceil N \cdot \frac{\theta - 0}{1 - 0} \right\rceil = \lceil N\theta \rceil.$$

The terms in this series have the same form as the ones in the earlier series, and they are all positive. So if this sum is to be smaller than the earlier one, it must extend over fewer terms. Thus $\Pr(Z_\infty \geq \theta) < \Pr(Z_G \geq \theta)$ if and only if the lower limit on the summation index is higher in the Z_∞ case than in the Z_G case, that is,

$$\lceil N\theta \rceil > \left\lceil N \cdot \frac{\theta - y_1}{y_2 - y_1} \right\rceil. \qquad (*)$$

A sufficient condition for this is

$$\theta \geq \frac{\theta - y_1}{y_2 - y_1} + \frac{1}{N}.$$

Note that this cannot possibly be satisfied if $N = 1$, since $\theta \in (0, 1)$. Suppose now $\theta > (\theta - y_1)/(y_2 - y_1)$, which is so whenever $\theta < y_1/(1 - y_2 + y_1)$. Then we can pick N large enough so that

$$\theta - \frac{\theta - y_1}{y_2 - y_1} > \frac{1}{N},$$

in which case the condition on the index is satisfied. Thus we have shown that if

$$\theta < \frac{y_1}{1 - y_2 + y_1},$$

then ρ_X is a gid.

To make this discussion even more concrete, here is a numerical example. For Y_G, we take a binomial rv that attains $y_1 = 3/8$ with probability $1 - p = 4/5$, and $y_2 = 7/8$ with probability $p = 1/5$. Thus Y_∞ is a binomial that attains 0 with probability $4/5$, and 1 with probability $1/5$. Then we take $\theta = 1/2$ and set $N = 3$, so that

$$Z_G = \sum_{i=1}^{3} Y_G/3 \quad \text{and} \quad Z_\infty = \sum_{i=1}^{3} Y_\infty/3.$$

Note that these choices satisfy $(*)$.

i	$\binom{N}{i}$	$(1-p)^{N-i}p^i$	Value of Z_G	Value of Z_∞	Probability of Attaining
0	1	64/125	9/24	0/3	64/125
1	3	16/125	13/24	1/3	48/125
2	3	4/125	17/24	2/3	12/125
3	1	1/125	21/24	3/3	1/125

$\Pr(Z_G \geq \theta)$	$\Pr(Z_\infty \geq \theta)$
61/125	13/125

Table 1: Behavior of the Sum of Three Binomial Random Variables

The values that Z_G and Z_∞ can attain, and the probabilities of attaining them, are straightforward to calculate. The discrete "pdfs" of these two rvs appear in Figure 8, and their behavior is summarized in Table 1. Note that of the values that Z_G can attain, only $z = 9/24$ lies below θ. But of the values that Z_∞ can attain, both $z = 0$ and $z = 1/3$ fall below θ. This shift is responsible for the drop in the tail probability.

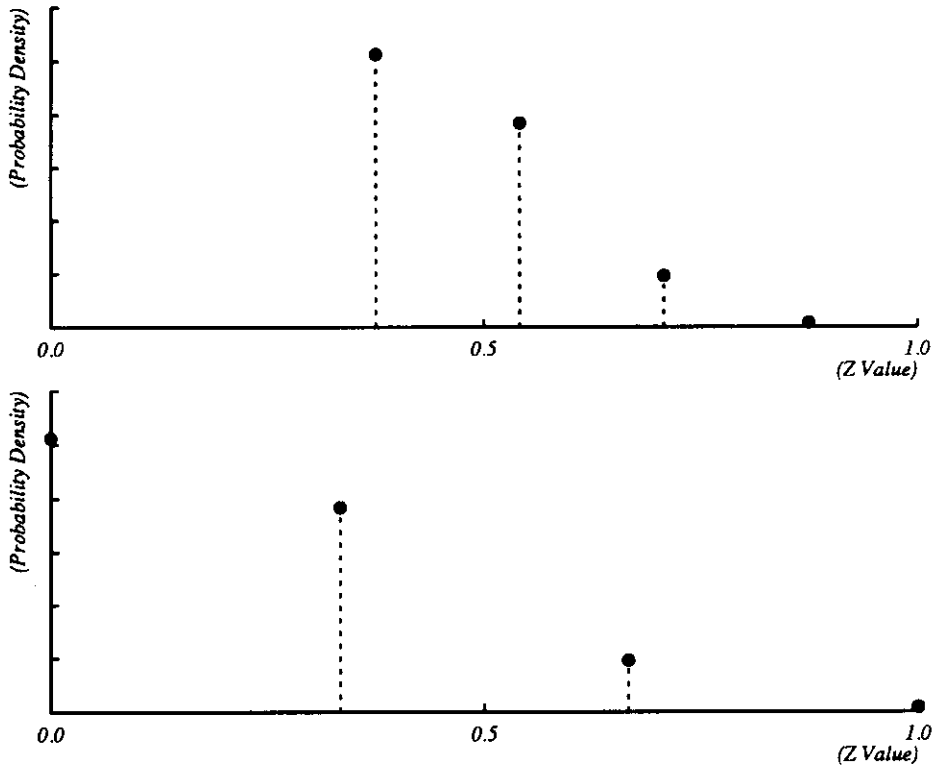

Figure 8: "Pdfs" of Z_G and Z_∞. These graphs depict the atoms of $Z_G = \sum_{i=1}^{3} Y_G/3$ (top) and $Z_\infty = \sum_{i=1}^{3} Y_\infty/3$ (bottom). Note how increasing the gain moves one of the atoms across the threshold $\theta = 1/2$.

4.2.3 Statement of the Theorem

Now we turn to formulating the Ensemble Performance Theorem. To do this, we need to elaborate on two ideas we introduced in the preceding paragraphs. First, we extend the notion of a rate number to cover the signal present case. This is just a matter of turning things around and looking at the opposite tail of the pdf. That is, the tail probability $\Pr(Z_{GS} \leq \theta)$ behaves like ${\gamma_{GS}}^N$ as N increases, where γ_{GS} lies in $[0, 1]$. Here γ_{GS} is defined by

$$\gamma_{GS} = \lim_{N \to \infty} \Pr(Z_{GS} \leq \theta)^{1/N}.$$

Likewise, we define $\gamma_{\infty S} = \lim_{G \to \infty} \gamma_{GS}$.

Second, we need to discuss the limiting distributions of Y_{GA} and Y_{GS}, as G increases without bound. It is intuitively clear that both these random variables tend to binomials on the values 0 and 1 as $G \to \infty$, and we will prove this in Section 6, Lemma 6. We write the limiting rvs as $Y_{\infty A}$ and $Y_{\infty S}$. Now we define the quantities

$$A^+ = \int_0^\infty \rho_{X_A} \quad \text{and} \quad S^+ = \int_0^\infty \rho_{X_S}.$$

As we demonstrate below, the probability that $Y_{\infty A}$ takes the value 1 is A^+, and the probability that $Y_{\infty S}$ takes the value 1 is S^+. It follows at once that $\mu(Y_{\infty A}) = A^+$ and $\mu(Y_{\infty S}) = S^+$.

It is straightforward to determine the rate number for a binomial random variable; we do this in Section 6, just after Lemma 10. For now we just state the results for $Y_{\infty A}$ and $Y_{\infty S}$. Providing $\mu(Y_{\infty A}) < \theta < \mu(Y_{\infty S})$, we have

$$\gamma(\theta, Y_{\infty A}) = \left(\frac{A^+}{\theta}\right)^\theta \left(\frac{1 - A^+}{1 - \theta}\right)^{1-\theta} \quad \text{and} \quad \gamma(\theta, Y_{\infty S}) = \left(\frac{S^+}{\theta}\right)^\theta \left(\frac{1 - S^+}{1 - \theta}\right)^{1-\theta}.$$

Why have we not written $\gamma_{\infty A}$ and $\gamma_{\infty S}$ on the lefthand side of these two equalities? This question goes to the heart of our proof. Recall that $\gamma_{\infty A}$ is defined as $\lim_{G \to \infty} \gamma_{GA}$, and likewise for $\gamma_{\infty S}$. But note that γ_{GA} is itself defined in terms of a limiting process—specifically, $\gamma_{GA} = \lim_{N \to \infty} \Pr(Z_{GA} \geq \theta)^{1/N}$. Thus to write $\gamma_{\infty A} = \gamma(\theta, Y_{\infty A})$ is to assert that

$$\lim_{G \to \infty} \left(\lim_{N \to \infty} \Pr(Z_{GA} \geq \theta)^{1/N} \right) = \lim_{N \to \infty} \left(\lim_{G \to \infty} \Pr(Z_{GA} \geq \theta) \right)^{1/N};$$

in other words, that the order of taking limits can be interchanged.

In fact, this is true—otherwise we would not have chosen so suggestive a notation. But proving it is quite difficult. For the moment we will take it as given, and proceed with our exposition. We are now ready to state the second main result of this paper.

> **Theorem (Ensemble Performance)** *Consider a multi-unit receiver operating at gain G on a non-trivial signal detection task $\mathcal{T}$. Let Θ be the set of real numbers θ satisfying any one of the following conditions*

- $\gamma(\theta, Y_{\infty A}) < \gamma(\theta, Y_{GA})$ *and* $\gamma(\theta, Y_{\infty S}) < \gamma(\theta, Y_{GS})$
- $\gamma(\theta, Y_{\infty S}) < \gamma(\theta, Y_{GS})$ *and* $\gamma(\theta, Y_{\infty A}) < \gamma(\theta, Y_{GS})$
- $\gamma(\theta, Y_{\infty A}) < \gamma(\theta, Y_{GA})$ *and* $\gamma(\theta, Y_{\infty S}) < \gamma(\theta, Y_{GA})$.

Then for each $\theta \in \Theta$, there is a $G' > G$ such that for sufficiently large N, $E_{G'}(\theta) > E_G(\theta)$. In particular, if $\theta^\star_G \in \Theta$, then $E^\star_{G'} > E^\star_G$. In other words, increasing the gain improves the optimal performance.

We will now attempt to explain the reasoning at work here, though a complete understanding requires tackling the proof.

Let's return to the performance function for a multi-unit network. $E^\star_G$, the optimal performance at gain G, is given by

$$E^\star_G = E_G(\theta^\star_G) = \lambda + \alpha \int_{\theta^\star_G}^1 \rho_{Z_{GS}} - \beta \int_{\theta^\star_G}^1 \rho_{Z_{GA}}.$$

We want to show that by increasing the gain to some new value G', we can improve upon this number. Note that the first integral equals $1 - \Pr(Z_{GS} \leq \theta^\star_G)$, and the second equals $\Pr(Z_{GA} \geq \theta^\star_G)$. Hence this expression can be written in terms of tail probabilities as

$$E^\star_G = \lambda + \alpha - \alpha \cdot \Pr(Z_{GS} \leq \theta^\star_G) - \beta \cdot \Pr(Z_{GA} \geq \theta^\star_G),$$

where the first tail is the chance of a miss, and the second is the chance of a false alarm.

Now α, β and λ are determined as before by the signal detection task $\mathcal{T}$. Thus to have any hope of getting better performance, we must somehow favorably alter the values of the tail probabilities. Since both α and β are positive, any change that simultaneously reduces both tails will improve performance. This will certainly happen when we increase the gain if both ρ_{X_S} and ρ_{X_A} are gids for the threshold $\theta^\star_G$. But it turns out that even when increasing the gain causes one tail to go up, if it does not go up too quickly, the overall performance can still improve.

Hence the heart of the problem is to determine when, if ever, any of these conditions are fulfilled. We will go at this in a slightly backward way. Starting from the optimal threshold $\theta^\star_G$ for a fixed gain G, we imagine increasing the gain without bound, and ask if $E_\infty(\theta^\star_G)$ exceeds $E_G(\theta^\star_G)$. If so, then we know there is some finite gain $G' > G$ such that $E_{G'}(\theta^\star_G) > E_G(\theta^\star_G)$, and hence that $E^\star_{G'}$, the optimal performance at G', exceeds $E^\star_G$, the optimal performance at G. From this we will try to work back-to-front and *deduce* sufficient conditions to make the optimal performance go up with increased gain.

Thus we are led to consider the expression

$$\Delta_\infty = E_\infty(\theta) - E_G(\theta)$$

for fixed θ, and ask under what conditions $\Delta_\infty > 0$. By simple arithmetic,

$$\Delta_\infty = \alpha \cdot \{\Pr(Z_{GS} \leq \theta) - \Pr(Z_{\infty S} \leq \theta)\} - \beta \cdot \{\Pr(Z_{\infty A} \geq \theta) - \Pr(Z_{GA} \geq \theta)\},$$

where α and β are determined as before by the payoffs and the prior probabilities.

Now the tail probabilities in this expression all vary with N, since each rv Z is defined as $\sum_{i=1}^{N} Y/N$ for the corresponding Y. As N gets big, they behave like

$$\Pr(Z_{GS} \leq \theta) \sim {\gamma_{GS}}^N \qquad \Pr(Z_{\infty S} \leq \theta) \sim {\gamma_{\infty S}}^N$$
$$\Pr(Z_{GA} \geq \theta) \sim {\gamma_{GA}}^N \qquad \Pr(Z_{\infty A} \geq \theta) \sim {\gamma_{\infty A}}^N.$$

(These equations express asymptotic relations that are not strictly true; the correct relations are of the slightly different form $\Pr^{1/N} \sim \gamma$.) Thus for large N,

$$\Delta_\infty \sim \alpha \left({\gamma_{GS}}^N - {\gamma_{\infty S}}^N\right) - \beta \left({\gamma_{\infty A}}^N - {\gamma_{GA}}^N\right).$$

There are now various cases to be considered, depending upon the directions of the inequalities $\gamma_{GS} \lessgtr \gamma_{\infty S}$ and $\gamma_{GA} \lessgtr \gamma_{\infty A}$.

Suppose for the moment that $\gamma_{GS} > \gamma_{\infty S}$. Then $\Delta_\infty > 0$ iff

$$\frac{\alpha}{\beta} > \frac{{\gamma_{\infty A}}^N - {\gamma_{GA}}^N}{{\gamma_{GS}}^N - {\gamma_{\infty S}}^N}. \tag{†}$$

Now if $\gamma_{\infty A} < \gamma_{GA}$, then ${\gamma_{\infty A}}^N - {\gamma_{GA}}^N < 0$ for all N, and since $\alpha, \beta > 0$, the inequality holds trivially. Otherwise, we have $\gamma_{\infty A} \geq \gamma_{GA}$, and (†) can be rewritten

$$\frac{\alpha}{\beta} > \left(\left(\frac{\gamma_{\infty A}}{\gamma_{GS}}\right)^N - \left(\frac{\gamma_{GA}}{\gamma_{GS}}\right)^N\right) \times \frac{1}{1 - \left(\gamma_{\infty S}/\gamma_{GS}\right)^N}. \tag{‡}$$

It follows by a simple limit argument that if $\gamma_{GS} > \gamma_{\infty A}$, from which $\gamma_{GS} > \gamma_{GA}$ as well, we can always find N sufficiently large to make this inequality true, *no matter what the values of α and β.*

This is a good point to take a look at the meaning of the inequalities among the rate numbers. Consider the conditions $\gamma_{GS} > \gamma_{\infty S}$ and $\gamma_{GA} > \gamma_{\infty A}$. As we argued on page 21, these respectively imply that ρ_{X_S} and ρ_{X_A} are gids. This shows that the first condition in the theorem implies improved performance with increasing gain.

More searchingly, consider the inequalities $\gamma_{GS} > \gamma_{\infty A}$ and $\gamma_{\infty A} \geq \gamma_{GA}$ that we have just treated. Since $\gamma_{GS} > \gamma_{\infty A}$, we know that for sufficiently large N, the quantity $(\gamma_{\infty A}/\gamma_{GS})^N$ can be made arbitrarily small. Thus even though $\gamma_{\infty A} \geq \gamma_{GA}$ means that ρ_{X_A} is *not* a gid, the condition $\gamma_{GS} > \gamma_{\infty A}$ ensures us that the right hand side of (‡) still falls to 0 with increasing N. Hence as the gain increases, even though the number of false alarms rises, the increase is small enough that for some fixed large N, the overall performance still goes up. This is the formal basis for our earlier comment that the ensemble effect can appear even when the number of false alarms rises with increasing gain, providing it does not go up too far.

Furthermore, by the definition of a limit, the inequalities $\gamma_{GS} > \gamma_{\infty S}$ and $\gamma_{GS} > \gamma_{\infty A}$ imply that we can find some *finite* G' such that $\gamma_{GS} > \gamma_{G'S}$ and $\gamma_{GS} > \gamma_{G'A}$, and

therefore make all these arguments go through for some sufficiently large N. That is, we have established that if

$$\gamma_{GS} > \gamma_{\infty S} \quad \text{and} \quad \gamma_{GA} > \gamma_{\infty A}$$

then for some finite $G' > G$ and some large, fixed but finite N, we have $E_{G'}(\theta) > E_G(\theta)$. This justifies the second condition in the theorem. The third condition, $\gamma_{GA} > \gamma_{\infty A}$ and $\gamma_{GA} > \gamma_{\infty S}$, is argued in the same way. Hence when any one of the three conditions holds, if θ happens to be θ^*_G, then $E_{G'}(\theta^*_G) > E_G(\theta^*_G) = E^*_G$. But $E^*_{G'} = E_{G'}(\theta^*_{G'}) \geq E_{G'}(\theta^*_G)$, so we have $E^*_{G'} > E^*_G$ as claimed. This completes our discussion of the proof.

Next we fulfill the promise, made at the start of this section, to show how the two conditions on X_S and X_A—that they are neither identical nor separable—follow from the hypotheses of the theorem. First suppose that they are identical; we will show this implies Θ is empty. For given any θ, either $\gamma_{\infty S}$ or $\gamma_{\infty A}$ is 1, since $\mu(Y_{\infty S}) = \mu(Y_{\infty A})$ belongs to either $[0, \theta]$ or $[\theta, 1]$. But the rate number for any rv satisfies $\gamma \leq 1$. Hence for any θ, at least one conjunct of each of the theorem's three conditions must fail, so Θ is empty.

On the other hand, suppose X_S and X_A are separable, say by the value x^*. We will show that $\theta^*_G \notin \Theta$. For if x^* separates X_S and X_A, then for each G, $f_G(x^*)$ separates Y_{GS} and Y_{GA}, and hence also Z_{GS} and Z_{GA}. Thus the threshold $\theta^*_G = f_G(x^*)$ gives perfect performance, and therefore is optimal for the gain G. Now observe that if θ^*_G separates Z_{GS} and Z_{GA}, then $\gamma_{GA} = 0$ and $\gamma_{GS} = 0$. But the rate number for any rv satisfies $\gamma \geq 0$. Hence for θ^*_G, none of the inequalities can hold, so $\theta^*_G \notin \Theta$.

4.2.4 Meaning of the Theorem, and a Numerical Example

Now for a few words about what the theorem means. The performance improvement is not achieved by increasing the number of units. As we mentioned before, we are *not* claiming that taking more units improves signal detection performance. Such a claim, while true, follows at once from the well-known result that taking more statistically independent samples reduces the expected error [9]. Our result is that with fixed N, providing the conditions of the theorem are fulfilled, we can improve the performance of the network *just by increasing the gain*. It is not even necessary to change the operating threshold from θ^*_G to $\theta^*_{G'}$, though doing so could improve the performance still further.

As in the case of the Constant Optimal Performance Theorem, there are no conditions on the activation family $\{f_G\}$. But in the case of the Ensemble Performance Theorem, this should be emended to "no *explicit* conditions on the activation family." For the shape of f_G plays a major role in determining the distributions of $Y_{GA} = f_G(X_A)$ and $Y_{GS} = f_G(X_S)$, and hence also in determining the rate numbers $\gamma(\theta, Y_{GA})$ and $\gamma(\theta, Y_{GS})$. Comparison of these quantities with the limiting values $\gamma_{\infty A}$ and $\gamma_{\infty S}$—which *are* independent of the particular activation family—then determines whether or not the ensemble effect appears at the given threshold.

But this does not represent a shortcoming of the theorem. Our definition of an activation family was deliberately unrestrictive, and it admits families of truly wild functions. It is entirely plausible that the shape of some given activation function f_G will influence whether or not the ensemble effect appears in this instance. For instance, f_G must not already be so close to the limiting step function u_0 that there is no room for improvement. Indeed, it is somewhat surprising that all the required information about f_G in relation to X_S and X_A can be bundled up in just four numbers—γ_{GS}, γ_{GA}, $\gamma_{\infty S}$ and $\gamma_{\infty A}$.

It must be said, however, that the current theorem does leave something to be desired. We would like the result to hold *uniformly*, at fixed G' and N, for all $\theta \in \Theta$. But the order of quantification in the statement of the theorem is

$$\forall\theta \in \Theta \quad \exists G' \quad \exists N.$$

Hence the increased gain G' and the size N of the multi-unit receiver required to see the effect can vary with the particular value of the threshold θ. This produces certain difficulties in applying the theorem. Consider some X_S, X_A and f_G. As we increase N to a number sufficiently large to ensure the appearance of the effect, in general $\theta^\star_G$ will change, and may in fact no longer lie in Θ!

Actually, the situation is not quite as bad as it seems. Suppose we are trying to determine conditions under which a multi-unit receiver will exhibit the ensemble effect on inputs X_S and X_A. The set Θ of admissible thresholds, though dependent upon the input rvs and f_G, does not depend at all upon α, β or N. So for any given gain G and threshold $\theta^\star_G$, we first compute the rate numbers and ensure that $\theta^\star_G \in \Theta$. Next we find G' large enough to bring $\gamma_{G'S}$ and $\gamma_{G'A}$ close to their limiting values, and then take N large enough for the effect to appear.

All we need now is a signal detection task such that the optimal threshold is actually $\theta^\star_G$. But this is easy to arrange; we just adjust α and β so that $\theta^\star_G$ is actually a solution of

$$\alpha \cdot \rho_{Z_{GS}}(\theta) = \beta \cdot \rho_{Z_{GA}}(\theta),$$

which determines the optimal threshold. (Consult Figure 6 to see how changes in α and β can be used to bring $\theta^\star_G$ to any value we desire.) We can then work backward from α and β to determine the payoffs and prior probabilities.

A stronger version of the theorem would reverse the order of the quantifiers, to

$$\exists N \quad \exists G' \quad \forall\theta \in \Theta.$$

Proving such a result would probably require more stringent conditions on the thresholds admitted to Θ, and in this sense would be narrower than our current formulation. But it would have the advantage of applying, for fixed N and G', to all the admissible thresholds.

The obstacle to proving such a version is that we have no information about how rapidly the tails $\Pr(Z_{GA} \geq \theta)$ and $\Pr(Z_{GS} \leq \theta)$ converge to $\gamma(\theta, Y_{GA})^N$ and $\gamma(\theta, Y_{GS})^N$ as N increases. Without this knowledge, we can say only that for *some* sufficiently

large N, they are close enough that we can ignore any differences, and thereby draw the desired conclusion.

We now provide a numerical example of the effect. We have already seen half of this example—the gid treated in Table 1 and Figure 8. We use this as the signal absent distribution. For the signal present distribution, we reflect the top graph of Figure 8 through a vertical line passing through the value .5 on the Z axis. Taking appropriate values for the payoffs and prior probabilities P_S and P_A, we obtain $\alpha = \beta = 1$. By a symmetry argument, $\theta^* = 1/2$. Then for low gain, we have

$$\begin{aligned} E_G^\star - \lambda &= \alpha \cdot \Pr(Z_{GS} \geq \theta^*) - \beta \cdot \Pr(Z_{GA} \geq \theta^*) \\ &= 64/125 - 61/125 \\ &= 3/125, \end{aligned}$$

and at high gain

$$\begin{aligned} E_G^\star - \lambda &= 112/125 - 13/125 \\ &= 99/125. \end{aligned}$$

We report the value of $E_G^\star - \lambda$ because λ is a constant depending only upon $\mathcal{T}$, and we are concerned with the gain-varying portion of the performance. Thus, this is an example of the ensemble effect. Figure 9 shows how Z_{GS} and Z_{GA} change with increasing gain, and yield the performance improvement.

Unfortunately, this example is not biologically plausible. First, the underlying input rvs X_S and X_A are both binomials, so they contain atoms. But this is not a serious problem, since there are atomless distributions that give the same numerical results. The real difficulty is that X_S and X_A have entirely different structures. They are not even close to being translates of one another. It is difficult to explain why the noise distribution, which is presumably additive at the network inputs, should depend upon the state of the world.

We have searched for a numerical example under the following conditions, which we consider biologically more plausible. We took each input to be a Gaussian rv, centered at x_S or x_A. That is, $X_S = x_S + W$, and $X_A = x_A + W$, where W is a zero-mean Gaussian rv of variance independent of the state of the world. For the activation family, we used the biased logistics, with bias -1, so

$$f_G(x) = \frac{1}{1 + e^{-(Gx-1)}}.$$

Then we experimented with the gain, the size of the network, and the parameters of the signal detection task in an attempt to find an instance of the effect.

It was not difficult to find a situation where both ρ_{X_S} and ρ_{X_A} were gain improvable densities, that is,

$$\gamma(\theta, Y_{\infty S}) < \gamma(\theta, Y_{GS}) \quad \text{and} \quad \gamma(\theta, Y_{\infty A}) < \gamma(\theta, Y_{GA}).$$

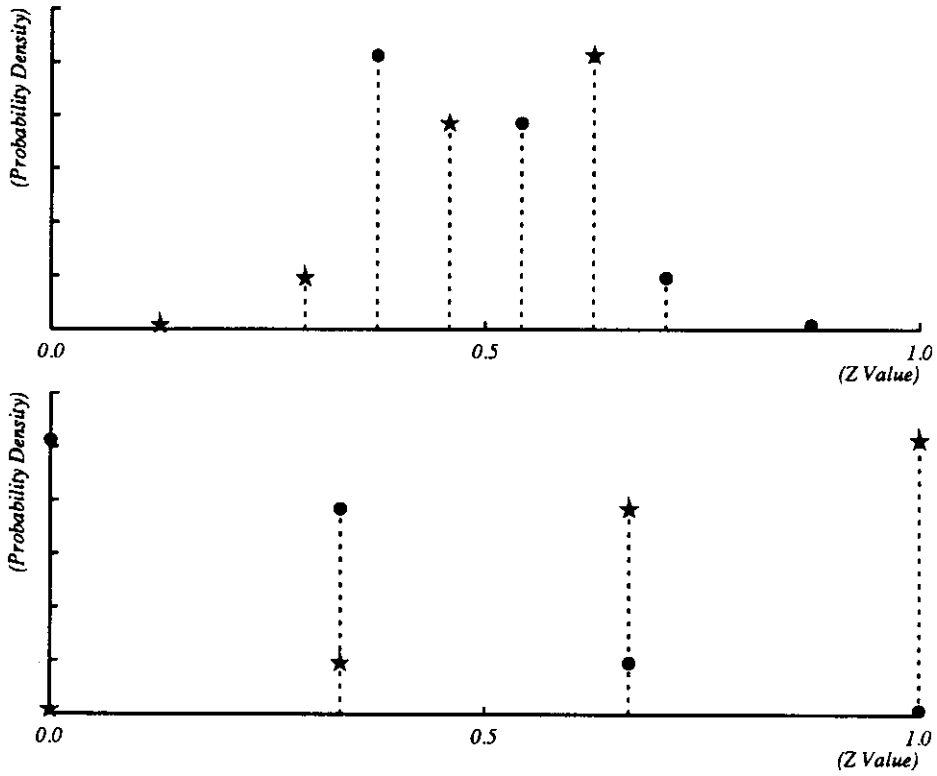

Figure 9: Ensemble Effect Example. These graphs depict the atoms of both the signal-absent rv Z_{GA}, marked with a $\bullet$, and the signal-present rv Z_{GS}, marked with a $\star$. The top graph shows them for some finite gain, the bottom graph shows them in the limit as $G \to \infty$. By symmetry, the optimal threshold θ^* equals $1/2$ in both graphs. Note how increasing the gain moves one of the signal-absent atoms below the threshold, and one of the signal-present atoms above the threshold, simultaneously reducing both the probability of a miss and the probability of a false alarm.

However, when we fixed N and explicitly computed the performance at optimal threshold for two different values of the gain, we obtained inconclusive results. We saw both small performance improvements and drops, on the order of .01%, for networks containing from 2 to 16 units. Since these results lie within the error range of our numerical integration procedure, they cannot be regarded as instances of the effect.

We have devoted some thought to these figures. We are not disturbed by the performance drops, since the theorem makes no prediction for small N. But when is N big enough? It appears that the tail probabilities converge slowly to their asymptotic form γ^N, and this means N must take substantially larger values than we have considered. Because we do not know a way to compute the required integrals with sufficient precision, we have not examined such high values of N.

Here is an example of the magnitude of N required. In one case we studied, we obtained the following rate numbers

G	γ_{GS}	γ_{GA}
1	.67367	.99867
∞	.64210	.98638

Hence both Y_{GS} and Y_{GA} are gids. For the argument behind the theorem to apply, $\Pr(Z_{GS} \leq \theta)$ and $\Pr(Z_{GA} \geq \theta)$ must be well-approximated by ${\gamma_{GS}}^N$ and ${\gamma_{GA}}^N$. But for the numbers in the table above, ${\gamma_{GA}}^N$ does not drop below .5 until $N \approx 512$.

These insights have led us to believe that the influence of the ensemble effect is small, at least in this situation. The theorem does not come into play, so to speak, until the network is reasonably large. By then almost all of the probability mass is concentrated around the two means. It follows that there is very little mass left in the tails, upon which the gain-varying portion of the performance depends. Thus even though increasing the gain may reduce both the miss and false alarm probabilities, these reductions, and therefore the performance improvement, are likely to be small.

4.3 The Chain Performance Theorem

In this section we state and explain the Chain Performance Theorem. As the name suggests, this result applies to a chain network, which is the third model in Figure 1. The theorem gives sufficient conditions for improved performance, with increasing gain, in the presence of additive noise at the output of a single-unit network. Not too surprisingly, we call this the *chain effect*.

After the sometimes difficult going of the previous section, this discussion will be easier to follow. This is because the intuition behind the effect is easy to grasp. We proceed to develop this intuition, in the following steps. First, we briefly review the chain network, with special attention to the role of noise and its sources. We also explain why we use the term "chain." Next we develop the basic intuition behind the theorem, by considering an example where there is *no* noise in the input to the network. This is biologically implausible, but conceptually useful. Then we state the theorem, and follow it, as usual, with some discussion of its meaning.

Refering to Figure 1, we review the operation of a chain. This is a single-unit network, but with the unit's activation y summed with a noise term v to yield the final output z. This added term is described by a random variable V, which we will refer to as the *output noise*. We write ρ_V for its pdf (when it exists), and in an abuse of notation, $V()$ for its distribution function.

We place no restrictions on the mean or variance of the output noise. This means that the final output rv Z is no longer concentrated on $(0, 1)$. It is easy to compute ρ_Z; this is just the convolution of ρ_Y and ρ_V.

In what follows, we make two key assumptions about the output noise. We assume that it arises from brain activity that is uncorrelated with the signal detection task at hand, and that its distribution is independent of the gain. These assumptions are evident in our notation for the output noise rv, which is an unadorned V. When we write X_S and X_A for the input rvs, the subscripts distinguish the signal-present and signal-absent cases. We make no such distinction for V, since by assumption it is independent of the stimulus. Likewise, V has no G subscript because it is gain-independent. We will say more about both these assumptions in the critique at the end of this document.

It is important to see that there is another reason why V bears no G subscript, arising not from any assumption, but from the structure of the network we are modelling. The activation rvs Y_{GS} and Y_{GA} are labelled this way because their distributions are determined by f_G, as $f_G(X_S)$ and $f_G(X_A)$. However, f_G has no direct influence upon V. Noise that is present in the inputs, which is captured in the rvs X_S and X_A, gets transformed by the activation function, whereas output noise does not. This is not an assumption, but an insight into a fundamental difference between the influence of gain variation upon noise arising at two different places in the network.

Proceeding as before, it is now straightforward to develop an expression for the performance function E. The network's output at gain G in the presence of signal is $Z_{GS} = Y_{GS} + V$, and in the absence of signal it is $Z_{GA} = Y_{GA} + V$. Hence

$$E(\theta) = \lambda + \alpha \int_\theta^\infty \rho_{Z_{GS}} - \beta \int_\theta^\infty \rho_{Z_{GA}}.$$

Note that both integrals now extend over all the reals above θ, since the output is no longer restricted to $(0, 1)$.

This is an appropriate spot to discuss our use of the word "chain." Let us imagine that the output of one neuron serves as the input to another. The random variable V models the noise that may be aggregated with the first neuron's output, in the dendritic tree of the post-synaptic neuron. But our figure does not show a unit corresponding to the second neuron, nor have we included it in the model.

This is because it is mathematically superfluous, as we shall now argue. We claim that under suitable conditions, increasing the gain improves the distinguishability of the rvs Z_{GS} and Z_{GA}, which describe the signal-present and signal-absent inputs to the second neuron. But by an argument identical to the one used to establish the Constant Optimal Performance Theorem, it is possible to show that the optimal performance achievable on the output of the second neuron is precisely the optimal performance

achievable on Z_{GS} and Z_{GA}. Hence it suffices, for our purposes, to show that E^*_G, as defined on Z_{GS} and Z_{GA}, goes up with increasing gain. This will then establish an improvement, with increasing gain, in the optimal performance at the final output of a chain of two neurons.

Now we explain how the effect arises. Suppose for the moment that the rvs X_S and X_A always took the values x_S and x_A respectively, with $x_A < 0 < x_S$. Then the rvs Y_{GS} and Y_{GA} would take the values $y_{GS} = f_G(x_S)$ and $y_{GA} = f_G(x_A)$. Since $\{f_G\}$ is an activation family, we know that a suitable increase in gain would cause y_{GS} to increase, and y_{GA} to decrease.

For the sake of illustration, let us now suppose that V were normally distributed, with mean 0 and variance σ^2. Then clearly $Z_{GS} = y_{GS} + V$ would be normally distributed with mean y_{GS}, and $Z_{GA} = y_{GA}+V$ would be normally distributed with mean y_{GA}. Thus the effect of increasing the gain would be to slide the entire distribution of Z_{GS} to higher activation values, leaving its shape unchanged, and likewise to slide Z_{GA} to lower values. This effect is illustrated in Figure 10. Hence for any fixed threshold θ, simultaneously $\Pr(Z_{GS} \geq \theta)$ would *rise*, and $\Pr(Z_{GA} \geq \theta)$ would *fall*, with increasing gain.

The first of these is the probability of a hit, and the second is the probability of a false alarm. Since

$$E(\theta) = \lambda + \alpha \cdot \Pr(Z_{GS} \geq \theta) - \beta \cdot \Pr(Z_{GA} \geq \theta),$$

this establishes that for *any* threshold θ, we have $E_{G'}(\theta) \geq E_G(\theta)$ for $G' > G$. The inequality is strict if G' is sufficiently larger than G. In particular, this holds for θ^*_G, so $E^*_{G'} \geq E_{G'}(\theta^*_G) > E_G(\theta^*_G) = E^*_G$, or more simply $E^*_{G'} > E^*_G$, which is the desired result.

While this argument gives a strong intuition about the effect, the situation is not quite as simple as Figure 10 suggests. In general, the input rvs overlap, in the manner of Figure 11. And by virtue of the Constant Optimal Performance Theorem, we know that even though an increase in gain may drive apart the means of the output rvs Y_{GS} and Y_{GA}, there are compensating changes in $\rho_{Y_{GS}}$ and $\rho_{Y_{GA}}$ that keep the performance at optimal threshold constant.

Nevertheless, as careful inspection of Figure 11 bears out, under fairly general conditions the probabilities $\Pr(Z_{GS} \geq \theta)$ and $\Pr(Z_{GA} \geq \theta)$ behave as the intuition just developed leads us to expect, and the performance rises with increasing gain. This is formalized in the following theorem, which gives sufficient conditions for the appearance of the chain effect.

Theorem (Chain Performance) *Consider a chain operating at threshold θ and gain G on a signal detection task T. Let*

$$A^+ = \int_0^\infty \rho_{X_A} \quad \textit{and} \quad S^+ = \int_0^\infty \rho_{X_S},$$

and let V() be the distribution function of the output noise. Then provid-

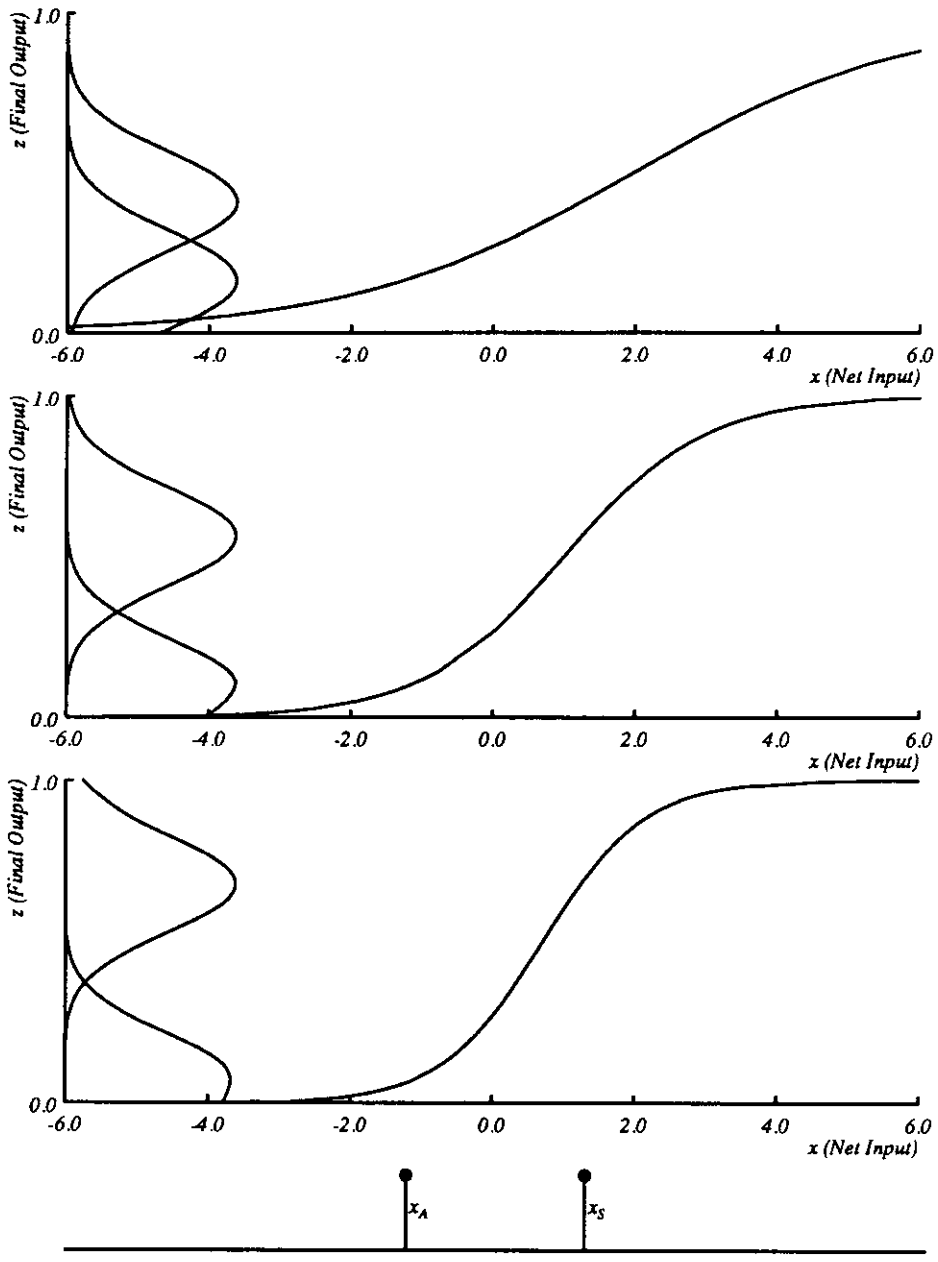

Figure 10: Dependence of Chain Output Pdfs Upon Gain. These graphs use the same conventions as Figure 3. They depict the sliding of the output pdfs as the gain moves through the values 0.5, 1.0 and 1.4 (top to bottom). The vertical lines at the bottom represent the "pdfs" of X_A and X_S.

ing

$$\left(1 - F_{Y_{GA}}(\theta)\right) \cdot \frac{V(\theta) - V(0)}{V(\theta) - V(\theta - 1)} > A^+$$

and

$$S^+ > \frac{V(\theta) - V(0) \cdot F_{Y_{GS}}(\theta)}{V(\theta) - V(\theta - 1)},$$

there exists $G' > G$ *such that* $E_{G'}(\theta) > E_G(\theta)$.

The proof of this theorem is quite simple. We write down the inequalities

$$\Pr(Z_{GA} \geq \theta) > \Pr(Z_{\infty A} \geq \theta) \quad \text{and} \quad \Pr(Z_{\infty S} \geq \theta) > \Pr(Z_{GS} \geq \theta).$$

The first of these states that in the limit as the gain increases without bound, the chance of a false alarm goes down. The second says that in the same limit, the chance of a hit goes up. The hypotheses of the theorem are sufficient conditions that the statements are simultaneously true, and this gives us the desired result. The details, which consist of bounding the convolution integrals that define $F_{Z_{GS}}$ and $F_{Z_{GA}}$, appear in Section 7 below.

The X_S and X_A "pdfs" in Figure 10 have atoms and are therefore inadmissible; we used them because they make it easy to understand how the effect arises. They are also non-overlapping. But the effect also appears when the input rvs can be represented by true pdfs, even if these pdfs overlap. In Figure 11 we proceed to exhibit such a case. We use the familiar input rvs and gain values of Figure 4; the output noise rv V is a zero-mean Gaussian. The numerical parameters used for this example are as follows

α	β	$\mu(X_S)$	$\sigma(X_S)$	$\mu(X_A)$	$\sigma(X_A)$	$\mu(V)$	$\sigma(V)$
1	1	1.25	1	−1.25	1	0	.15

and here is the performance at optimal threshold for each gain value.

G	$\theta_G^\star$	$E_G^\star - \lambda$
0.5	.299	.495
1.0	.328	.662
1.4	.344	.711

At the end of Section 4.1, where we pointed out a deficiency in theories of catecholamine effects based upon the signal-to-noise ratio (SNR), we also said that these explanations were not without merit. We touch on this again now.

The heart of the matter is that the chain effect arises because increasing the gain helps drown out the noise along the communication pathway in a chain of neurons. In a chain of two neurons, this is effectively the same as increasing the SNR at the input to the second one. We present a more complete discussion in Section 8 below.

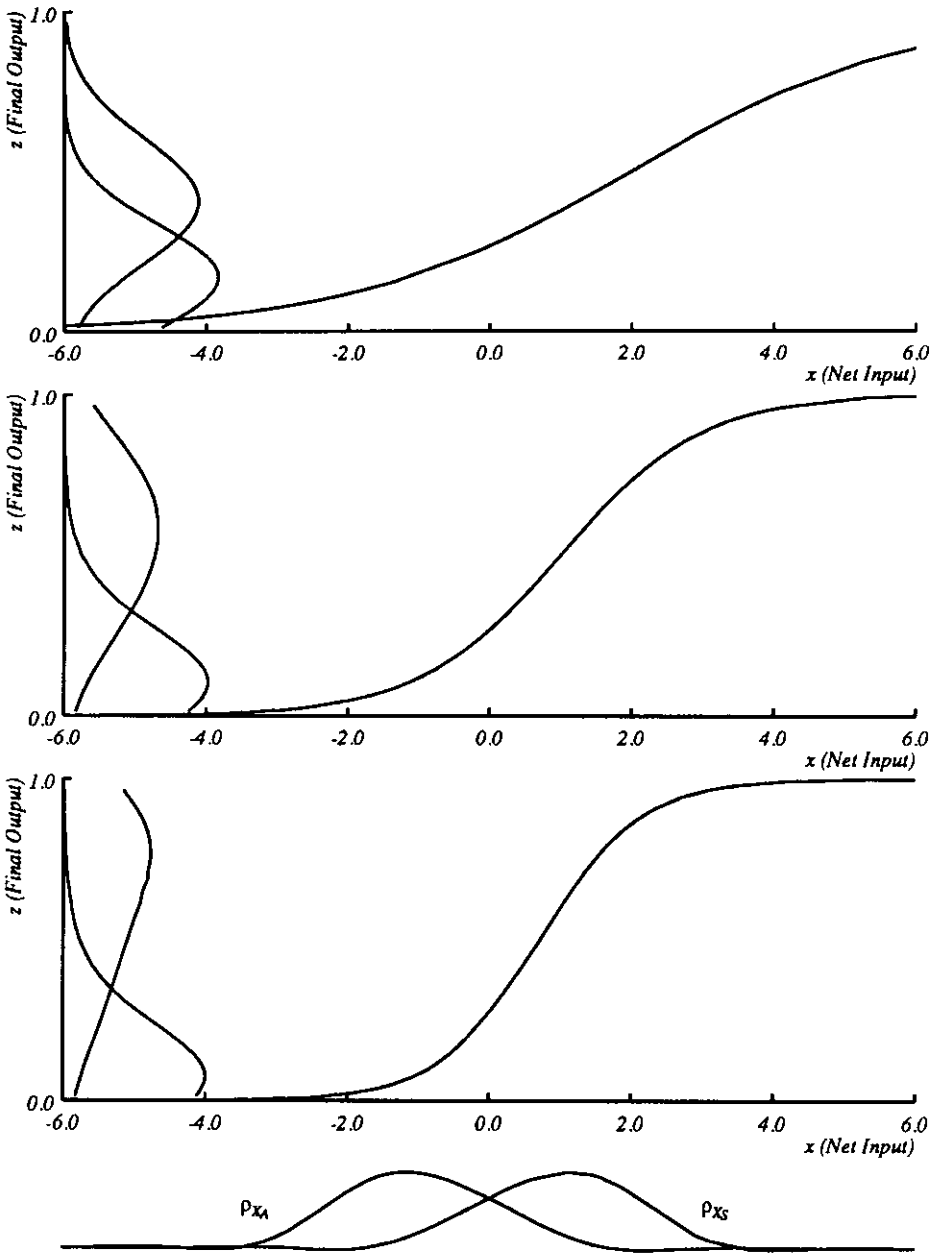

Figure 11: Dependence of Chain Output Pdfs Upon Gain. These graphs use the same conventions and input pdfs as Figure 3. They depict the output pdfs, in the presence of additive Gaussian noise, as the gain moves through the values 0.5, 1.0 and 1.4 (top to bottom).

5 Constant Optimal Performance Theorem: Proof

In this section we prove the Constant Optimal Performance Theorem. We start with an overview of the argument. The first step is to define a new function, $\mathcal{E}_G(x)$, that expresses the performance $E_G(\theta)$ as a function of $x = f_G^{-1}(\theta)$, the inverse image of the threshold. Since for some thresholds θ there may not even be an x such that $\theta = f_G(x)$, this may seem impossible, but as we will see, there is a way around this difficulty. Next we show that for any two gain values G and G', the functions $\mathcal{E}_G$ and $\mathcal{E}_{G'}$ are identical. This result lies at the heart of the proof, since once it has been demonstrated, we know that any performance achievable at one gain G is achievable at any other gain G'. We argue exactly this point and thereby establish the theorem.

Now for the proof. First, a little notation. We will write $\Re$ for the real numbers, and $\Re^\dagger$ for the extended reals, obtained by adjoining $-\infty$ and ∞ to $\Re$. Now let $\mathcal{T}$ be a signal detection task, and consider a single-unit network with activation family $\{f_G\}$ and performance function E_G. Recall that E_G is defined as

$$E_G(\theta) = \lambda + \alpha \int_\theta^1 \rho_{Y_{GS}} - \beta \int_\theta^1 \rho_{Y_{GA}},$$

where $Y_{GS} = f_G(X_S)$ and $Y_{GA} = f_G(X_A)$. We define the function $\mathcal{E}_G$ on $\Re$ by

$$\mathcal{E}_G = E_G \circ f_G$$

and enlarge its domain to $\Re^\dagger$ by adjoining the ordered pairs $\langle -\infty, \lambda + \alpha - \beta \rangle$ and $\langle \infty, \lambda \rangle$ to the function.

Lemma 1 *For any two gain values G and G', the functions $\mathcal{E}_G$ and $\mathcal{E}_{G'}$ are identical.*

Proof: To show that $\mathcal{E}_G$ and $\mathcal{E}_{G'}$ are the same function, we must show they are equal for every $x \in \Re^\dagger$. By construction they are equal at $-\infty$ and at ∞, so it suffices to consider only $x \in \Re$. Now

$$\mathcal{E}_G(x) = E_G(f_G(x)) = \lambda + \alpha \int_{f_G(x)}^1 \rho_{Y_{GS}} - \beta \int_{f_G(x)}^1 \rho_{Y_{GA}}.$$

By definition,

$$\begin{aligned} \int_{f_G(x)}^1 \rho_{Y_{GS}} &= \Pr(Y_{GS} \geq f_G(x)) \\ &= \Pr(f_G(X_S) \geq f_G(x)) \quad \text{by definition of } Y_{GS} \\ &= \Pr(X_S \geq x) \quad \text{since } f_G \text{ is strictly increasing} \\ &= \int_x^\infty \rho_{X_S}. \end{aligned}$$

A similar argument shows that

$$\int_{f_G(x)}^1 \rho_{Y_{GA}} = \int_x^\infty \rho_{X_A}.$$

Hence

$$\mathcal{E}_G(x) = \lambda + \alpha \int_x^\infty \rho_{X_S} - \beta \int_x^\infty \rho_{X_A}.$$

But the expression on the right is independent of the gain, and since the same argument applies to $\mathcal{E}_{G'}$, this establishes our claim. ∎

In view of this result, we will confine our attention to a single function $\mathcal{E}$ on $\Re$ given by

$$\mathcal{E}(x) = \lambda + \alpha \int_x^\infty \rho_{X_S} - \beta \int_x^\infty \rho_{X_A},$$

with its domain enlarged to $\Re^\dagger$ as before. Intuitively, $\mathcal{E}$ expresses the performance in a gain-independent way. We now show that $\mathcal{E}$ attains the same values as E_G, and only those values.

Lemma 2 *For any G, the range of E_G equals the range of $\mathcal{E}$.*

Proof: The range of E_G is the set of all values $E_G(\theta)$ where θ varies over $[0, 1]$. Since $\mathcal{E} = E_G \circ f_G$ for any G, clearly *range* $\mathcal{E} \subseteq$ *range* E_G. So it suffices to prove the reverse inclusion.

We need to show that for every $\theta \in [0, 1]$, we have $E_G(\theta) \in$ *range* $\mathcal{E}$. That is, we must exhibit an $x \in \Re^\dagger$ such that $E_G(\theta) = \mathcal{E}(x)$. So suppose $\theta \in [0, 1]$ is given, and let $L = \{x \in \Re \mid f_G(x) \geq \theta\}$. Define $x_\theta \in \Re^\dagger$ by

$$x_\theta = \begin{cases} \infty & \text{if } L \text{ is empty,} \\ -\infty & \text{if } L \text{ is unbounded below, and} \\ \inf L & \text{otherwise.} \end{cases}$$

We claim $\mathcal{E}(x_\theta) = E_G(\theta)$. We proceed to examine the three possibilities.

First suppose L is empty. Then $f_G(x) < \theta$ for all $x \in \Re$. This means

$$\Pr(Y_{GS} \geq \theta) = \Pr(f_G(X_S) \geq \theta) = 0,$$

and likewise for Y_{GA}. Hence $E_G(\theta) = \lambda = \mathcal{E}(x_\theta)$.

Next suppose L is unbounded below. This implies that $f_G(x) \geq \theta$ for all $x \in \Re$. For suppose that $f_G(x_1) < \theta$ for some $x_1 \in \Re$. Since L is unbounded below, there is some $x_0 < x_1$ for which $f_G(x_0) \geq \theta$. But f_G is increasing, so $f_G(x_1) \geq f_G(x_0) \geq \theta$, contrary to our supposition.

Now since $f_G(x) \geq \theta$ for all $x \in \Re$, we have

$$\Pr(Y_{GS} \geq \theta) = \Pr(f_G(X_S) \geq \theta) = 1,$$

and likewise for Y_{GA}. Hence $E_G(\theta) = \lambda + \alpha - \beta = \mathcal{E}(x_\theta)$.

Finally, suppose $x_\theta = \inf L$. Now L is the set of all points such that $f_G(x) \geq \theta$; that is, $x \in L \Leftrightarrow f_G(x) \geq \theta$. Thus

$$\Pr(Y_{GS} \geq \theta) = \Pr(f_G(X_S) \geq \theta) = \Pr(X_S \in L).$$

Now L is either the set (x_θ, ∞) or $[x_\theta, \infty)$. But

$$\int_{(x_\theta,\infty)} \rho_{X_S} = \int_{[x_\theta,\infty)} \rho_{X_S} = \int_{x_\theta}^{\infty} \rho_{X_S}$$

since ρ_{X_S} is Lebesgue integrable, so adjoining or deleting x_θ from the domain of integration does not change the integral. Thus

$$\Pr(Y_{GS} \geq \theta) = \int_{x_\theta}^{\infty} \rho_{X_S},$$

and a similar argument applies to Y_{GA}, so we have

$$\begin{aligned} E_G(\theta) &= \lambda + \alpha \cdot \Pr(Y_{GS} \geq \theta) - \beta \cdot \Pr(Y_{GA} \geq \theta) \\ &= \lambda + \alpha \int_{x_\theta}^{\infty} \rho_{X_S} - \beta \int_{x_\theta}^{\infty} \rho_{X_A} \\ &= \mathcal{E}(x_\theta) \end{aligned}$$

as desired. ∎

All the hard work has now been done, and we are ready to prove the theorem.

Theorem 1 (Constant Optimal Performance) *Let $E_G^\star$ be the performance at optimal threshold of a single-unit network on the signal detection task T, and let $\{f_G\}$ be the unit's activation family. Then $E_G^\star$ is a constant, independent of the gain G.*

Proof: Let G and G' be arbitrary gain values. Define the function $\mathcal{E}$ on $\Re^\dagger$ as above. By Lemma 2,

$$E_G^\star = \sup \mathit{range}\ E_G = \sup \mathit{range}\ \mathcal{E} = \sup \mathit{range}\ E_{G'} = E_{G'}^\star.$$

But G and G' were arbitrary. Hence the optimal performance $E_G^\star$ is the same for each value of the gain. ∎

6 Ensemble Performance Theorem: Proof

In this section we prove the Ensemble Performance Theorem. The proof is a straightforward limit calculation, which closely parallels the discussion that follows the original statement of the theorem, in Section 4.2.3. However, to put this argument on a firm mathematical foundation, we need to establish two supporting results. First, we must prove the existence of the limit

$$\lim_{N\to\infty} \Pr(Z \geq \theta)^{1/N},$$

which defines the rate number $\gamma(\theta, Y)$. Second, we must develop techniques for reasoning about the tail probabilities $\Pr(Z_{GA} \geq \theta)$ and $\Pr(Z_{GS} \leq \theta)$ as $G \to \infty$. The

latter result itself consists of two parts: determining the limiting behavior of the rate numbers γ_{GS} and γ_{GA}, and showing how to reduce questions about tail probabilities to questions about rate numbers.

Unfortunately, the road to establishing these results is somewhat long and arduous. To help the reader follow it, we now provide an overview of the whole development.

We begin with a random variable X, and establish that $Y_G = f_G(X)$ and $Y_\infty = u_0(X)$ really are random variables as well. That is, they define probability measures on $\Re$. This allows us to exploit the standard integral convergence theorems.

Next we introduce the function M_Y, which is known as the *moment generating function* (mgf) of Y. (Now and hereafter Y stands either for $Y_\infty = u_0(X)$, or any $Y_G = f_G(X)$.) This function turns out to be a key instrument in computing rate numbers and reasoning about them, and we will study it in some detail. In particular, we prove that as $G \to \infty$, we have $M_{Y_G} \to M_{Y_\infty}$ uniformly on any compact set. It follows directly from this that $F_{Y_G} \to F_{Y_\infty}$ as $G \to \infty$, which we claimed earlier but did not prove. But the real reason we establish this result is that it is key in showing that $\gamma_G \to \gamma_\infty$, which is essential to proving the theorem.

To establish this last result, we proceed in a slightly roundabout way. First we introduce the *rate function*, I_Y, and show how it is related to the mgf. Then we apply Cramér's Theorem to establish that $\lim_{N\to\infty} \Pr(Z \geq \theta)^{1/N}$ exists and equals $e^{-I_Y(\theta)}$. In a single stroke this shows both that the rate number is well-defined, and gives us a way to compute it. We immediately apply this result to determine the rate number for a binomial rv.

Then in a sequence of lemmas we prove that $\gamma_G \to \gamma_\infty$ as $G \to \infty$. This is important because the very next step is to show that certain questions about tail probabilities can be reduced to questions about their rate numbers. This is the "convergence trick" we spoke of earlier. Thus if we know the limiting values of the rate numbers, these can be used to draw conclusions about the limiting behavior of the tail probabilities, and hence of the performance itself. With the requisite tools finally in hand, we then proceed to prove the theorem. This concludes the overview.

We start by establishing some notation. Through most of this discussion, we will generally not distinguish the signal present and signal absent cases. For the present we are concerned only with the general properties of rate numbers and moment generating functions.

In what follows, X is a random variable on $\Re$, which should be thought of as standing for either X_S or X_A—we dispense temporarily with the S and A subscripts. X is assumed to have a Lebesgue-integrable pdf ρ_X. We write F_X for its distribution function (df), defined by $F_X(\xi) = \Pr(X \leq \xi)$. As before, $\{f_G\}$ is an activation family, and Y_G is the rv defined by $Y_G = f_G(X)$. We write F_{Y_G} for the df of Y_G, and ρ_{Y_G} for its pdf, when this exists. We also define Y_∞ as $u_0(X)$, where u_0 is the step function at 0, the limiting form of f_G.

Lemma 3 *Y_∞ is a random variable, and Y_G is a random variable for each G.*

Proof: By [2, Theorem 3.1.4], it suffices to show that u_0, and each f_G, are Borel measurable. Since these are all increasing functions on $\Re$, this follows at once. ∎

Clearly, Y_∞ is a binomial rv with df F_{Y_∞} given by

$$F_{Y_\infty}(y) = \begin{cases} 0 & \text{for} \quad y < 0 \\ 1-p & \text{for} \quad 0 \le y < 1 \\ p & \text{for} \quad 1 \le y \end{cases}$$

where $p = \int_0^\infty \rho_X$.

Now we introduce the moment generating function of Y. This is a real-valued function, written $M_Y(\xi)$, defined for all $\xi \in \Re$ by

$$M_Y(\xi) = \int_{-\infty}^{\infty} e^{\xi y} \rho_Y(y)\, dy = \int_0^1 e^{\xi y} \rho_Y(y)\, dy.$$

The last equality follows because Y is concentrated on $[0, 1]$. Note that if Y does not have a pdf, this can equally well be written as the Lebesgue-Stieltjes integral

$$M_Y(\xi) = \int_0^1 e^{\xi y}\, dF_Y(y).$$

Since we have placed no bound upon ξ in this expression, the reader may have a queasy feeling that all sorts of delicate issues of convergence are being swept under the rug. However, we will now show that we are in the fortunate position that this integral transform always exists for the rvs in question. At the same time, we prove several well-known properties of the mgf.

Lemma 4 *Let Y be one of the rvs $Y_G = f_G(X)$ for some G, or $Y_\infty = u_0(X)$. Then (1) the moment generating function $M_Y(\xi)$ exists for all $\xi \in \Re$, (2) M_Y has derivatives of all orders everywhere on $\Re$, (3) $M_Y(\xi)$ is always strictly greater than zero, and (4) $\log M_Y(\xi)$ is a convex function of ξ.*

Proof: We have

$$M_Y(\xi) = \int_0^1 e^{\xi y}\, dF_Y(y).$$

Since $e^{\xi y}$ is continuous as a function of y for each ξ, it is bounded on $[0, 1]$, so the existence of the integral follows at once. To prove (2), let $M_Y^{(k)}$ denote the kth derivative of the mgf; a similar argument establishes the existence of

$$M_Y^{(k)}(\xi) = \int_0^1 y^k e^{\xi y}\, dF_Y(y)$$

for every positive integer k. The differentiation under the integral sign is justified because $y^k e^{\xi y}$ is continuous, hence bounded on $[0, 1]$, for each k.

To prove (3), note that by Jensen's inequality [10, Proposition 5.17], for any $\xi \in \Re$,

$$M_Y(\xi) = \int_0^1 e^{\xi y}\, dF_Y(y) \;\ge\; e^{\xi \mu(Y)} \;>\; 0.$$

Finally, for any $\alpha \in (0,1)$, and all $\xi_1, \xi_2 \in \Re$,

$$M_Y(\alpha\xi_1 + (1-\alpha)\xi_2) \leq M_Y(\xi_1)^{\alpha} M_Y(\xi_2)^{1-\alpha}$$

by Hölder's inequality [10, Theorem 6.2]. Taking log of both sides gives us (4). ∎

Next we show that it is meaningful to talk about the limiting behavior of Y_G as $G \to \infty$. We will show that $M_{Y_G} \to M_{Y_\infty}$ as $G \to \infty$, and deduce from this that $F_{Y_G} \to F_{Y_\infty}$. But we will actually establish a much stronger result, namely that M_{Y_G} converges uniformly to M_{Y_∞} on any compact set. This is by far the most searching proof in the paper, and we will put it to good use later on.

Lemma 5 (Uniform Convergence of MGFs) *For any compact $D \subseteq \Re$, we have $M_{Y_G} \to M_{Y_\infty}$, and $\log M_{Y_G} \to \log M_{Y_\infty}$, both uniformly on D, as $G \to \infty$.*

Proof: First note that it suffices to prove that $M_{Y_G} \to M_{Y_\infty}$ uniformly on D. The uniform convergence of $\log M_{Y_G}$ to $\log M_{Y_\infty}$ follows from this because M_{Y_∞} is bounded uniformly away from zero on D, and log is uniformly continuous on any compact domain.

We proceed to establish the uniform convergence of M_{Y_G} to M_{Y_∞} on D. Let $\{G_n\}$ be any increasing sequence of gain values such that $G_n \to \infty$ as $n \to \infty$. This gives us a sequence of functions $\{M_{Y_{G_n}}\}$. Our strategy is to decompose each $M_{Y_{G_n}}$ as the sum of two continuous functions, L_n and U_n, such that each function sequence $\{L_n\}$ and $\{U_n\}$ is isotone, and where $L_n \to 1-p$ and $U_n \to pe^{\xi}$ as $n \to \infty$. Then by Dini's Theorem [7, Theorem 6.11], both $\{L_n\}$ and $\{U_n\}$ converge uniformly on D, so $M_{Y_{G_n}} \to (1-p) + pe^{\xi}$ uniformly on D. By direct evaluation of the integral, $M_{Y_\infty}(\xi) = (1-p) + pe^{\xi}$. Thus $M_{Y_{G_n}} \to M_{Y_\infty}$ uniformly on D, and since this holds for every increasing divergent sequence $\{G_n\}$, we have the desired result.

So it is enough to construct the sequences $\{L_n\}$ and $\{U_n\}$. First we show how to reduce any M_{Y_G} to an expectation involving f_G and X. Let us write $\varepsilon[R]$ for the expectation of a random variable R. Then we have

$$M_{Y_G}(\xi) = \int_0^1 e^{\xi y}\, dF_{Y_G}(y) = \varepsilon[e^{\xi Y_G}] = \varepsilon[e^{\xi f_G(X)}] = \int_{-\infty}^{\infty} e^{\xi f_G(x)}\, \rho_X(x)\, dx,$$

where the last equality follows from the definition of expectation, and the assumption that X has a Lebesgue-integrable pdf ρ_X.

Fix $\xi \in D$, and define

$$g_n(x) = \begin{cases} e^{\xi f_{G_n}(x)} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \quad \text{and} \quad h_n(x) = \begin{cases} e^{\xi f_{G_n}(x)} & \text{for } x \leq 0 \\ 0 & \text{for } x > 0 \end{cases}$$

We will presently show that each g_n and h_n is integrable, so we may define

$$U_n(\xi) = \int_{-\infty}^{\infty} g_n(x)\, dx \quad \text{and} \quad L_n(\xi) = \int_{-\infty}^{\infty} h_n(x)\, dx.$$

Clearly $M_{Y_{G_n}} = L_n + U_n$.

Now for every n, $e^{\xi f_{G_n}(x)}$ is measurable as a function of x, and ρ_X is integrable hence measurable, so each g_n and h_n is measurable. Also,

$$|h_n(x)|, |g_n(x)| \leq \max\{1, e^{\xi}\} \cdot \rho_X(x),$$

where the righthand side integrable for every ξ. This proves that each g_n and h_n is integrable. It also proves that we may apply the Dominated Convergence Theorem to conclude $U_n \to pe^{\xi}$ and $L_n \to 1 - p$ pointwise as $n \to \infty$. A similar argument shows that each U_n and L_n is differentiable, hence continuous as a function of ξ.

It remains only to show that each sequence of functions is isotone. At this point, we separate D into $D^+ = D \cap [0, \infty)$ and $D^- = D \cap (-\infty, 0]$. Note that D^+ and D^- are compact. We will show that $\{U_n\}$ and $\{L_n\}$ are isotone on D^+ and D^- separately. This proves that $U_n \to pe^{\xi}$, and $L_n \to 1 - p$, uniformly on each of D^+ and D^-. But then clearly the convergence is uniform over all of D, and we have the desired result.

So we proceed to show that each sequence is isotone as required. First note that if $m > n$ then $G_m \geq G_n$, since the sequence of gains is increasing. Thus $f_{G_m} \geq f_{G_n}$ on $[0, \infty)$, and $f_{G_m} \leq f_{G_n}$ on $(-\infty, 0]$, since $\{f_G\}$ is an activation family. Now let $\xi \in D^+$. Then $\xi \geq 0$, so

$$\begin{aligned} e^{\xi f_{G_m}(x)} &\geq e^{\xi f_{G_n}(x)} && \text{for all } x \geq 0, \text{ so} \\ g_m(x) &\geq g_n(x) && \text{for all } x \in \Re, \text{ so} \\ U_m(\xi) &\geq U_n(\xi) && \text{for all } \xi \in D^+. \end{aligned}$$

Thus $\{U_n\}$ is an increasing sequence of functions on D^+. Likewise, $\{L_n\}$ is a decreasing sequence of functions on D^+. Hence both sequences converge uniformly on D^+. A similar argument applies to D^-. ∎

This immediately gives us

Lemma 6 *$F_{Y_G} \to F_{Y_\infty}$ as $G \to \infty$. In other words, Y_G converges to the binomial rv Y_∞ as $G \to \infty$.*

Proof: Let us write $\mathcal{L}_Y$ for the Laplace Transform of Y; clearly $\mathcal{L}_Y(\xi) = M_Y(-\xi)$, where $\xi \geq 0$. By the preceding lemma, $\lim_{G \to \infty} \mathcal{L}_{Y_G}(\xi) = \mathcal{L}_{Y_\infty}(\xi)$ pointwise for all $\xi \in \Re$. Hence by [4, Theorem XIII.1.2], $F_{Y_G} \to F_{Y_\infty}$ as $G \to \infty$. ∎

We now embark on a series of lemmas that will provide the connection between the mgf and the rate number. These begin with the following definition. Let Y be either Y_∞ or some Y_G, and fix a threshold θ. We define the function $H_\theta(\xi)$ for each $\xi \in \Re$ by

$$H_\theta(\xi) = \xi\theta - \log M_Y(\xi).$$

Without delay, we prove

Lemma 7 *For any Y_G or Y_∞, the function $H_\theta(\xi)$ exists for all $\xi \in \Re$ and all $\theta \in [0, 1]$, and $-H_\theta(\xi)$ is a convex function of ξ.*

Proof: Immediate from Lemma 4. ∎

We will shortly establish that $\log \gamma(\theta, Y) = -\sup_{\xi \in \Re} H_\theta(\xi)$. But before we do this, we need to come to grips with an issue that we have so far finessed. In our motivating discussion of $\gamma(\theta, Y)$, we defined this quantity by the equation

$$\gamma(\theta, Y) = \lim_{N \to \infty} \Pr(Z \geq \theta)^{1/N},$$

where $Z = \sum_{i=1}^{N} Y/N$. Hence if $\Pr(Y \geq \theta)$ vanishes, so does $\Pr(Z \geq \theta)$ for all N, and therefore $\gamma(\theta, Y) = 0$. But then $\log \gamma(\theta, Y)$ does not exist!

In fact, if $\Pr(Y \geq \theta)$ vanishes, then H_θ is unbounded above as a function of ξ, so $\sup_{\xi \in \Re} H_\theta(\xi)$ does not exist. This pathology warns us that before constructing any mathematical apparatus that uses $\sup_{\xi \in \Re} H_\theta(\xi)$, we had best determine just when this quantity exists. The following lemma serves the need.

Lemma 8 *Given $\theta > \mu(Y)$, suppose that $\Pr(Y \geq \theta)$ is non-zero. Then $\sup_{\xi \in \Re} H_\theta(\xi)$ exists, and equals $\sup_{\xi \geq 0} H_\theta(\xi)$.*

Proof: First we show that if $\theta > \mu(Y)$, then $H_\theta(\xi) \leq H_\theta(0)$ for all $\xi \leq 0$. By Jensen's inequality, $M_Y(\xi) \geq e^{\xi \mu(Y)}$ for any ξ, so $-\log M_Y(\xi) \leq -\xi \mu(Y)$. Hence for $\xi \leq 0$,

$$H_\theta(\xi) = \xi\theta - \log M_Y(\xi) \;\leq\; \xi\,(\theta - \mu(Y)) \;\leq\; 0 = H_\theta(0).$$

This shows that $\sup_{\xi \in \Re} H_\theta(\xi) = \sup_{\xi \geq 0} H_\theta(\xi)$, providing it exists. We now establish this. Let

$$R_\theta(\xi) = e^{-\xi\theta} M_Y(\xi).$$

By Lemma 4, we have $R_\theta(\xi) > 0$ for all ξ. Hence $\log R_\theta(\xi)$ exists for all ξ, and clearly $H_\theta = -\log R_\theta$. Since log is continuous and strictly increasing on $(0, \infty)$, we have

$$\sup_{\xi \geq 0} H_\theta(\xi) = -\log \inf_{\xi \geq 0} R_\theta(\xi).$$

This equation means that if either side exists, then so does the other, and they are equal.

We proceed to show that the infimum on the right-hand side exists and is non-zero. Observe that for any $\xi \geq 0$, we have $e^{\xi(y-\theta)} \geq 1$ for all $y \geq \theta$. Hence

$$R_\theta(\xi) = \int_0^1 e^{\xi(y-\theta)} dF_Y(y) \;\geq\; \int_\theta^1 e^{\xi(y-\theta)} dF_Y(y) \;\geq\; \int_\theta^1 dF_Y(y) \;=\; \Pr(Y \geq \theta).$$

Since this bound holds for all $\xi \geq 0$, and since $\Pr(Y \geq \theta)$ is non-zero, we have $\inf_{\xi \geq 0} R_\theta(\xi) > 0$. Thus $\sup_{\xi \geq 0} H_\theta(\xi)$ exists. ∎

This is a good place to pause and consider another issue, one that is not so much mathematical as pedagogical. The preceding lemma required that $\theta > \mu(Y)$. Though we have temporarily suppressed the separation of signal-present and signal-absent cases, no doubt it is clear—harkening back to our motivating discussion in

Section 4.2—that we have the signal-absent case in mind here. But an analogous result can be proved for the signal-present case. The requisite changes in the hypotheses and the conclusions are straightforward, and not particularly interesting—for instance, we require that $\theta < \mu(Y)$ and that $\Pr(Y \leq \theta)$ is non-zero, and we get $\sup_{\xi \in \Re} H_\theta(\xi) = \sup_{\xi \leq 0} H_\theta(\xi)$.

So from here on we will pursue the development in a one-sided fashion. We will continue the investigation for rvs that satisfy $\theta > \mu(Y)$ and $\Pr(Y \geq \theta) \neq 0$, and leave it to the reader to reverse the sense of the inequalities as required.

Now we provide the long-promised link between the moment generating function and the rate number. For each $\theta \in \Re$ we define

$$I_Y(\theta) = \sup_{\xi \in \Re} \{\xi\theta - \log M_Y(\xi)\}.$$

Evidently, $I_Y(\theta) = \sup_{\xi \in \Re} H_\theta(\xi)$. I_Y is called the *rate function* for Y, and we will now see how it is connected with the rate number.

Lemma 9 (Existence and Computation of Rate Numbers) *Let Y be Y_∞, or Y_G for some gain G. Define $Z = \sum_{i=1}^{N} Y/N$, let $\theta \in (0, 1)$ be given, and define the rate function I_Y as above. Then providing $\theta > \mu(Y)$, and $\Pr(Y \geq \theta)$ is non-zero,*

$$\gamma(\theta, Y) = e^{-I_Y(\theta)}.$$

Proof: By Cramér's Theorem [15, Theorem 3.8] and Lemma 4 the random variable Z satisfies the large deviation principle with rate function I_Y. Thus

$$\limsup_{N \to \infty} \frac{1}{N} \log \Pr(Z \geq \theta) \leq -\inf_{x \geq \theta} I_Y(x)$$

and

$$\liminf_{N \to \infty} \frac{1}{N} \log \Pr(Z > \theta) \geq -\inf_{x > \theta} I_Y(x).$$

By [15, Lemma 3.3], I_Y is convex and hence continuous, so

$$\inf_{x \geq \theta} I_Y(x) = \inf_{x > \theta} I_Y(x) = I_Y(\theta),$$

which exists by virtue of Lemma 8. Now trivially, $\Pr(Z > \theta) \leq \Pr(Z \geq \theta)$. Hence

$$\liminf_{N \to \infty} \frac{1}{N} \log \Pr(Z > \theta) \leq \liminf_{N \to \infty} \frac{1}{N} \log \Pr(Z \geq \theta).$$

So we have

$$-I_Y(\theta) \leq \liminf_{N \to \infty} \frac{1}{N} \log \Pr(Z \geq \theta) \leq \limsup_{N \to \infty} \frac{1}{N} \log \Pr(Z \geq \theta) \leq -I_Y(\theta),$$

and this establishes the validity of the limit equation

$$\lim_{N \to \infty} \frac{1}{N} \log \Pr(Z \geq \theta) = I_Y(\theta).$$

The final conclusion follows from this by the continuity of exp, and the definition of $\gamma(\theta, Y)$. ∎

This lemma establishes two key results. First, it demonstrates the existence of the limit that defines $\gamma(\theta, Y)$. Second, it gives us an analytic tool for computing $\gamma(\theta, Y)$ for any Y_G, and even for Y_∞, which has atoms at 0 and 1. To help us apply these results, we prove the obvious but handy

Lemma 10 *If ξ_0 satisfies $H'_\theta(\xi_0) = 0$, then $\sup_{\xi \in \Re} H_\theta(\xi) = H_\theta(\xi_0)$.*

Proof: Consider the function $C(\xi) = -H_\theta(\xi)$; clearly $C'(\xi_0) = 0$. By Lemma 4, C is convex and differentiable everywhere. By [10, Proposition 5.16], C' is an increasing function.

Suppose there were some ξ_1 such that $H_\theta(\xi_1) > H_\theta(\xi_0)$. Then we should have $C(\xi_0) > C(\xi_1)$; we now show this is impossible.

Consider the chord from $\langle \xi_0, C(\xi_0) \rangle$ to $\langle \xi_1, C(\xi_1) \rangle$. If $\xi_0 < \xi_1$, the slope of this chord is negative. Hence by the Mean Value Theorem, there is some $\xi > \xi_0$ with $C'(\xi) < 0$. But this cannot be, since C' is increasing, and $C'(\xi_0) = 0$. A similar argument shows we cannot have $H_\theta(\xi_1) > H_\theta(\xi_0)$ for any $\xi_1 < \xi_0$. Thus $H_\theta(\xi_0)$ is the maximum. ∎

It is now straightforward to find the rate function of the binomial rv Y_∞, and we proceed to sketch this computation. Suppose Y_∞ attains the value 1 with probability p, and 0 with probability $1 - p$. Then by direct evaluation of the integral,

$$M_{Y_\infty}(\xi) = (1 - p) + pe^{\xi}.$$

Now we maximize $H_\theta(\xi) = \xi\theta - \log M_{Y_\infty}(\xi)$ by solving

$$H'_\theta(\xi) = \theta - \frac{1}{M_{Y_\infty}(\xi)} M'_{Y_\infty}(\xi) = 0.$$

So we require

$$\theta M_{Y_\infty}(\xi) = M'_{Y_\infty}(\xi),$$

that is,

$$\theta(1 - p + pe^{\xi}) = pe^{\xi}.$$

This may be solved for

$$\xi_0 = \log\left(\frac{\theta}{1-\theta}\right)\left(\frac{1-p}{p}\right).$$

Note that this solution is unique. By substituting ξ_0 into H_θ, and performing some enjoyable simplifications, we obtain

$$I_{Y_\infty}(\theta) = \log\left(\frac{\theta}{p}\right)^{\theta}\left(\frac{1-\theta}{1-p}\right)^{1-\theta}$$

for $0 < p < \theta < 1$. This gives us at once

$$\gamma(\theta, Y_\infty) = \left(\frac{p}{\theta}\right)^\theta \left(\frac{1-p}{1-\theta}\right)^{1-\theta},$$

which justifies our unproven claims in Section 4.2 about $\gamma_{\infty S}$ and $\gamma_{\infty A}$.

The last step before proving the theorem itself is to establish the limiting behavior of $\gamma(\theta, Y_G)$. We will show that $\lim_{G\to\infty} \gamma(\theta, Y_G) = \gamma(\theta, Y_\infty)$. But as we mentioned previously, this is a rather deep result. For it asserts that we may interchange the order of limits in the expression

$$\lim_{G\to\infty} \lim_{N\to\infty} \left(F_{Y_G}^{*N}(\theta)\right)^{1/N}.$$

Lemma 6 established only the pointwise convergence of F_{Y_G} to F_{Y_∞}. This convergence is *not* uniform, since each F_{Y_G} is continuous, whereas F_{Y_∞} plainly is not.

Despite this pathology, it does happen that $\gamma(\theta, Y_G) \to \gamma(\theta, Y_\infty)$ as $G \to \infty$, and we will now go about demonstrating this. The argument is separated into two lemmas. To make them easier to state and prove, we introduce a slight variation on our earlier notation. Let us add a G subscript to H_θ, writing

$$H_{\theta G}(\xi) = \xi\theta - \log M_{Y_G}(\xi) \quad \text{and} \quad H_{\theta\infty}(\xi) = \xi\theta - \log M_{Y_\infty}(\xi).$$

These are exactly the same as H_θ taken for Y_G and Y_∞. We introduce the notation because we are now concerned with the limiting behavior as G increases, so we want to display the G dependence explicitly.

Now for a sketch of the argument. Suppose $H_{\theta\infty}$ attains its sup at ξ_∞, and $H_{\theta G}$ attains its sup at ξ_G. In the first lemma, we show that for sufficiently large G, the value ξ_G lies close to ξ_∞. This permits us to concentrate our attention on a compact domain D. In the second lemma, by an appeal to the uniform convergence of $\log M_{Y_G}$ to $\log M_{Y_\infty}$ on D, we show that $H_{\theta G}(\xi_G) \to H_{\theta\infty}(\xi_\infty)$. Since $\gamma(\theta, Y_G) = H_{\theta G}(\xi_G)$ and $\gamma(\theta, Y_\infty) = H_{\theta\infty}(\xi_\infty)$, this gives us the desired result.

Lemma 11 (Localization of Maxima) *Let $Y_G = f_G(X)$ and $Y_\infty = u_0(X)$, with $p = \int_0^\infty \rho_X$. Suppose θ and p satisfy $0 < p < \theta < 1$, and let $H_{\theta\infty}(\xi)$ attain its maximum on $\Re$ at ξ_∞. Then there exists a gain G_0, and a compact $D \subseteq \Re$ containing ξ_∞, such that for all $G > G_0$, the function $H_{\theta G}(\xi)$ attains its maximum on $\Re$ at some $\xi_G \in D$.*

Proof: Set $l = \xi_\infty - 1$ and $h = \xi_\infty + 1$, and let D be the interval $[l, h]$. Since $H_{\theta\infty}$ attains its maximum uniquely at ξ_∞, we have the strict inequalities $H_{\theta\infty}(l), H_{\theta\infty}(h) < H_{\theta\infty}(\xi_\infty)$. Let $\epsilon = H_{\theta\infty}(\xi_\infty) - \max\{H_{\theta\infty}(l), H_{\theta\infty}(h)\}$; clearly $\epsilon > 0$.

Since D is compact, by Lemma 5 $\log M_{Y_G} \to \log M_{Y_\infty}$, and hence also $H_{\theta G} \to H_{\theta\infty}$, both uniformly on D. So there exists G_0 such that if $G > G_0$, then $|H_{\theta G}(\xi) - H_{\theta\infty}(\xi)| < \epsilon/3$ for all $\xi \in D$.

Now fix any $G > G_0$, and consider the chord from $\langle l, H_{\theta G}(l)\rangle$ to $\langle \xi_\infty, H_{\theta G}(\xi_\infty)\rangle$. We know

$$H_{\theta G}(l) \le H_{\theta\infty}(l) + \epsilon/3 < H_{\theta\infty}(\xi_\infty) - \epsilon/3 \le H_{\theta G}(\xi_\infty).$$

Thus the slope of this chord,

$$\frac{H_{\theta G}(\xi_\infty) - H_{\theta G}(l)}{1},$$

is strictly positive. By Lemma 4, the function $H_{\theta G}$ has derivatives of all orders everywhere. Hence by the Mean Value Theorem, it has a positive derivative for some ξ in (l, ξ_∞). A similar argument shows that it has a negative derivative for some ξ in (ξ_∞, h). But $H_{\theta G}'$ is continuous, since $H_{\theta G}$ has derivatives of all orders. Hence $H_{\theta G}$ has a zero derivative, and by Lemma 10 attains its supremum, for some ξ in $(l, h) \subseteq D$. ∎

Lemma 12 (Rate Number Convergence) *Let $Y_G = f_G(X)$ and $Y_\infty = u_0(X)$, with $p = \int_0^\infty \rho_X$. Suppose θ and p satisfy $0 < p < \theta < 1$. Then $\lim_{G\to\infty} \gamma(\theta, Y_G) = \gamma(\theta, Y_\infty)$.*

Proof: Let $\epsilon > 0$ be given. By the preceding lemma, we can find a gain G_1 and a compact $D \subseteq \Re$ such that for all $G > G_1$, we have $\gamma(\theta, Y_G) = \sup_{\xi\in D} H_{\theta G}(\xi)$.

Since $H_{\theta G} \to H_{\theta\infty}$ uniformly on D, there exists G_2 such that if $G > G_2$, then $|H_{\theta G}(\xi) - H_{\theta\infty}(\xi)| < \epsilon$ for all $\xi \in D$. Thus for $G > G_2$, we have

$$|\sup_{\xi\in D} H_{\theta G}(\xi) - \sup_{\xi\in D} H_{\theta\infty}(\xi)| < \epsilon.$$

Now take $G_0 = \max\{G_1, G_2\}$. Then for all $G > G_0$, we have

$$|\gamma(\theta, Y_G) - \gamma(\theta, Y_\infty)| = |\sup_{\xi\in D} H_{\theta G}(\xi) - \sup_{\xi\in D} H_{\theta\infty}(\xi)| < \epsilon,$$

as desired. ∎

We will need one more mathematical tool, to evaluate limits of form $\mathcal{P}_1/\mathcal{P}_2$ as $N \to \infty$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ are tail probabilities. As we now demonstrate, such limits are determined by the rate numbers of the two tails. The proof exploits the same basic trick that is used to establish the root test for convergence of infinite sequences.

Lemma 13 (Root Convergence) *Consider rvs Y_1 and Y_2 with $\gamma(\theta, Y_1) < \gamma(\theta, Y_2)$. Let $Z_1 = \sum_{i=1}^N Y_1/N$ and $Z_2 = \sum_{i=1}^N Y_2/N$. Then*

$$\Pr(Z_1 \geq \theta)/\Pr(Z_2 \geq \theta) \to 0 \quad \text{as } N \to \infty.$$

Proof: Let us write $\mathcal{P}_1 = \Pr(Z_1 \geq \theta)$, $\mathcal{P}_2 = \Pr(Z_2 \geq \theta)$, and $\gamma_1 = \gamma(\theta, Y_1)$, $\gamma_2 = \gamma(\theta, Y_2)$. Then by Lemma 9, we have

$$\mathcal{P}_1^{1/N} \to \gamma_1 \quad \text{and} \quad \mathcal{P}_2^{1/N} \to \gamma_2 \qquad \text{as } N \to \infty.$$

Since both these limits exist, and since our assumption $\gamma_2 > \gamma_1$ implies $\gamma_2 > 0$, we have

$$(\mathcal{P}_1/\mathcal{P}_2)^{1/N} \to \gamma_1/\gamma_2 \quad \text{as } N \to \infty. \qquad (*)$$

Now $\gamma_1 < \gamma_2$, so $\gamma_1/\gamma_2 < 1$. Pick ρ satisfying $(\gamma_1/\gamma_2) < \rho < 1$. Then by $(\star)$ there exists N_1 such that for all $N > N_1$

$$\mathcal{P}_1/\mathcal{P}_2 < \rho^N.$$

Now let $\epsilon > 0$ be given, and pick N_2 such that $\rho^{N_2} < \epsilon$. Set $N = \max\{N_1, N_2\}$; then for any $N' > N$ we have $\mathcal{P}_1/\mathcal{P}_2 < \rho^{N'} < \rho^N < \epsilon$. ∎

There are three additional versions of this lemma, with the inequalities of Z_1 and Z_2 with respect to θ set in the other possible combinations. They are all argued in exactly the same way, and we will not bother to state or prove them.

We are now in a position to prove the Ensemble Performance Theorem. But the formal argument obscures, rather than illuminates, what is really going on. For this reason, we urge the reader to return to the conclusion of Section 4.2, and reread the informal discussion of the argument there before attacking the proof.

Theorem 2 (Ensemble Performance) *Consider a multi-unit receiver operating at gain G on a non-trivial signal detection task $\mathcal{T}$. Let Θ be the set of real numbers θ satisfying any one of the following conditions*

- $\gamma(\theta, Y_{\infty A}) < \gamma(\theta, Y_{GA})$ *and* $\gamma(\theta, Y_{\infty S}) < \gamma(\theta, Y_{GS})$
- $\gamma(\theta, Y_{\infty S}) < \gamma(\theta, Y_{GS})$ *and* $\gamma(\theta, Y_{\infty A}) < \gamma(\theta, Y_{GS})$
- $\gamma(\theta, Y_{\infty A}) < \gamma(\theta, Y_{GA})$ *and* $\gamma(\theta, Y_{\infty S}) < \gamma(\theta, Y_{GA})$.

Then for each $\theta \in \Theta$, there is a $G' > G$ such that for sufficiently large N, $E_{G'}(\theta) > E_G(\theta)$. In particular, if $\theta^\star_G \in \Theta$, then $E^\star_{G'} > E^\star_G$. In other words, increasing the gain improves the optimal performance.

Proof: First note that since $\mathcal{T}$ is non-trivial, $\alpha, \beta > 0$ throughout. Now recall that

$$\begin{aligned} E_G(\theta) &= \lambda + \alpha - \alpha \Pr(Z_{GS} \leq \theta) - \beta \Pr(Z_{GA} \geq \theta), \quad \text{and} \\ E_{G'}(\theta) &= \lambda + \alpha - \alpha \Pr(Z_{G'S} \leq \theta) - \beta \Pr(Z_{G'A} \geq \theta) \end{aligned}$$

To keep things uncluttered, we write

$$\begin{aligned} \mathcal{P}_{GS} &= \Pr(Z_{GS} \leq \theta) \qquad \mathcal{P}_{G'S} = \Pr(Z_{G'S} \leq \theta) \\ \mathcal{P}_{GA} &= \Pr(Z_{GA} \geq \theta) \qquad \mathcal{P}_{G'A} = \Pr(Z_{G'A} \geq \theta). \end{aligned}$$

Let $\Delta_{G'} = E_{G'}(\theta) - E_G(\theta)$. Then by simple arithmetic we have

$$\Delta_{G'} = \alpha\{\mathcal{P}_{GS} - \mathcal{P}_{G'S}\} - \beta\{\mathcal{P}_{G'A} - \mathcal{P}_{GA}\}.$$

We proceed to show that each of the three conditions in the theorem implies that for some finite G' and N, we shall have $\Delta_{G'} > 0$.

Case 1: $\gamma_{\infty A} < \gamma_{GA}$ and $\gamma_{\infty S} < \gamma_{GS}$.

By Lemma 12, we can find $G' > G$ such that $\gamma_{GS} > \gamma_{G'S}$ and $\gamma_{GA} > \gamma_{G'A}$. Hence by Lemma 13, we can find N_1 and N_2 such that

$$\begin{aligned} \mathcal{P}_{GS} &> \mathcal{P}_{G'S} \quad \text{for all } N > N_1, \text{ and} \\ \mathcal{P}_{GA} &> \mathcal{P}_{G'A} \quad \text{for all } N > N_2 \end{aligned}$$

Take $N_0 = \max\{N_1, N_2\}$, then for all $N > N_0$ we have

$$\begin{aligned} \Delta_{G'} &= \alpha\{\mathcal{P}_{GS} - \mathcal{P}_{G'S}\} - \beta\{\mathcal{P}_{G'A} - \mathcal{P}_{GA}\} \\ &= \alpha\{\mathcal{P}_{GS} - \mathcal{P}_{G'S}\} + \beta\{\mathcal{P}_{GA} - \mathcal{P}_{G'A}\} \\ &> 0 \end{aligned}$$

since α and β, and both expressions in braces, are all strictly positive.

Case 2: $\gamma_{\infty S} < \gamma_{GS}$ and $\gamma_{\infty A} < \gamma_{GS}$.

The preceding argument covered the case when $\gamma_{GA} > \gamma_{\infty A}$, so without loss of generality assume $\gamma_{GA} \leq \gamma_{\infty A}$. Thus we have $\gamma_{GS} > \gamma_{GA}$ as well. Once again,

$$\Delta_{G'} = \alpha\{\mathcal{P}_{GS} - \mathcal{P}_{G'S}\} - \beta\{\mathcal{P}_{G'A} - \mathcal{P}_{GA}\}.$$

By the same reasoning as before, we can find $G' > G$ such that $\gamma_{GS} > \gamma_{G'S}$, and $\gamma_{GS} > \gamma_{G'A}$. Likewise, there is an N_1 such that $\mathcal{P}_{GS} > \mathcal{P}_{G'S}$ and for all $N > N_1$. Hence for all such N, $\Delta_{G'} > 0$ iff

$$\frac{\alpha}{\beta} > \frac{\mathcal{P}_{G'A} - \mathcal{P}_{GA}}{\mathcal{P}_{GS} - \mathcal{P}_{G'S}} = \left(\frac{\mathcal{P}_{G'A}}{\mathcal{P}_{GS}} - \frac{\mathcal{P}_{GA}}{\mathcal{P}_{GS}}\right) \times \frac{1}{1 - \mathcal{P}_{G'S}/\mathcal{P}_{GS}}.$$

But by Lemma 13,

$$\frac{\mathcal{P}_{G'A}}{\mathcal{P}_{GS}} \to 0, \quad \frac{\mathcal{P}_{GA}}{\mathcal{P}_{GS}} \to 0 \quad \text{and} \quad 1 - \mathcal{P}_{G'S}/\mathcal{P}_{GS} \to 1$$

as $N \to \infty$. Hence

$$\left(\frac{\mathcal{P}_{G'A}}{\mathcal{P}_{GS}} - \frac{\mathcal{P}_{GA}}{\mathcal{P}_{GS}}\right) \times \frac{1}{1 - \mathcal{P}_{G'S}/\mathcal{P}_{GS}} \to 0$$

as $N \to \infty$. By the definition of a limit, this means that we can find some N_2 such that for all $N > N_2$

$$\frac{\alpha}{\beta} > \frac{\mathcal{P}_{G'A} - \mathcal{P}_{GA}}{\mathcal{P}_{GS} - \mathcal{P}_{G'S}}.$$

So taking $N_0 = \max\{N_1, N_2\}$, we have $\Delta_{G'} > 0$ for all $N > N_0$.

Case 3: $\gamma_{\infty A} < \gamma_{GA}$ and $\gamma_{\infty S} < \gamma_{GA}$.

This is just like case 2. ∎

7 Chain Performance Theorem: Proof

In this section we prove the Chain Performance Theorem. Since the proof involves no new concepts, our discussion will be brief.

Our aim is to establish that by increasing the gain in a chain network to a suitably large value, we can improve the performance. To do so we will find conditions ensuring that with a sufficient gain increase, both the probability of false alarm and the probability of a miss will fall. We accomplish this in three steps. First we find the limiting values of these probabilities as $G \to \infty$. Then we develop bounds on them at finite gain. Finally we write down inequalities between the bounds and the limiting values, which assert that in the limit both probabilities decline. These expressions become the hypotheses of the theorem, which then follows at once.

We recall our earlier notation. Suppressing the S and A subscripts for a moment, we write X for an input rv, $Y_G = f_G(X)$ for the activation rv, and V for the output noise rv. In an abuse of notation, we will also write just $V()$, instead of $F_V()$, for V's distribution function. Z_G, the rv of the final output, is defined as $Y_G + V$. We write Y_∞ for $u_0(X)$, and Z_∞ for $Y_\infty + V$. Throughout the discussion, our only requirement is that $\{f_G\}$ is an activation family, and that X is atomless.

Lemma 14 *Let f_G, X, Y_G, V and Z_G be as given, and let $X^+ = \Pr(X \geq 0)$. Then*

$$F_{Z_\infty}(\theta) = \lim_{G\to\infty} F_{Z_G}(\theta) = X^+ \cdot (V(\theta - 1) - V(\theta)) + V(\theta).$$

Proof: First we establish that $F_{Z_\infty}(\theta) = \lim_{G\to\infty} F_{Z_G}(\theta)$. Since $Z_\infty = Y_\infty + V$, by [4, Theorem V.4.2], we have $F_{Z_\infty} = F_{Y_\infty} * V$ and $F_{Z_G} = F_{Y_G} * V$. Here we have written $*$ for the convolution. Thus

$$\lim_{G\to\infty} F_{Z_G}(\theta) = \lim_{G\to\infty} \int_{-\infty}^{\infty} F_{Y_G}(\theta - \xi)\, dV(\xi).$$

By the Dominated Convergence Theorem for Lebesgue-Stieltjes integrals, and Lemma 6, we have

$$\lim_{G\to\infty} \int_{-\infty}^{\infty} F_{Y_G}(\theta - \xi)\, dV(\xi) = \int_{-\infty}^{\infty} F_{Y_\infty}(\theta - \xi)\, dV(\xi) = F_{Z_\infty}(\theta),$$

where the last equality follows from the definition of Z_∞. To get the expression for $F_{Z_\infty}(\theta)$, we perform the integration explicitly, and do a little arithmetic. ∎

Next we develop upper and lower bounds on the distribution function F_{Z_G} at fixed finite gain.

Lemma 15 *Let Y_G, Z_G and V be defined as above. Then*

$$F_{Y_G}(\theta) \cdot V(\theta) \;\leq\; F_{Z_G}(\theta) \;\leq\; F_{Y_G}(\theta) \cdot (V(\theta) - V(0)) + V(0).$$

Proof: We will establish both bounds on $F_{Z_G}(\theta)$ at the same time. First we write down the convolution integral for F_{Z_G}, and decompose it into three terms.

$$\begin{aligned} F_{Z_G}(\theta) &= (F_{Y_G} * V)(\theta) \\ &= \int_{-\infty}^{\infty} F_{Y_G}(\theta - \xi)\, dV(\xi) \\ &= \int_{-\infty}^{0} F_{Y_G}(\theta - \xi)\, dV(\xi) + \int_{0}^{\theta} F_{Y_G}(\theta - \xi)\, dV(\xi) + \int_{\theta}^{\infty} F_{Y_G}(\theta - \xi)\, dV(\xi). \end{aligned}$$

The last integral vanishes, because $F_{Y_G}(\theta - \xi) = 0$ for all $\xi \geq \theta$. Now we bound the two remaining integrals. The following two lines should be read from the center out in both directions.

$$F_{Y_G}(\theta) \cdot V(0) = F_{Y_G}(\theta) \int_{-\infty}^{0} dV(\xi) \leq \int_{-\infty}^{0} F_{Y_G}(\theta - \xi)\, dV(\xi) \leq \int_{-\infty}^{0} dV(\xi) = V(0),$$

and

$$0 = F_{Y_G}(0) \int_{0}^{\theta} dV(\xi) \leq \int_{0}^{\theta} F_{Y_G}(\theta - \xi)\, dV(\xi) \leq F_{Y_G}(\theta) \int_{0}^{\theta} dV(\xi) = F_{Y_G}(\theta) \cdot (V(\theta) - V(0)).$$

Adding these up, we get the desired conclusions. ∎

Now we proceed with the proof.

Theorem 3 (Chain Performance) *Consider a chain operating at threshold θ and gain G on a signal detection task $\mathcal{T}$. Let*

$$A^+ = \int_{0}^{\infty} \rho_{X_A} \quad \textit{and} \quad S^+ = \int_{0}^{\infty} \rho_{X_S},$$

and let $V()$ be the distribution function of the output noise. Then providing

$$\left(1 - F_{Y_{GA}}(\theta)\right) \cdot \frac{V(\theta) - V(0)}{V(\theta) - V(\theta - 1)} > A^+$$

and

$$S^+ > \frac{V(\theta) - V(0) \cdot F_{Y_{GS}}(\theta)}{V(\theta) - V(\theta - 1)},$$

there exists $G' > G$ such that $E_{G'}(\theta) > E_G(\theta)$.

Proof: Let

$$\Delta_\infty(\theta) = E_\infty(\theta) - E_G(\theta),$$

where $E_\infty(\theta) = \lim_{G' \to \infty} E_{G'}(\theta)$. Then it suffices to show that $\Delta_\infty(\theta) > 0$. By simple arithmetic

$$\Delta_\infty(\theta) = \alpha\{\Pr(Z_{GS} \leq \theta) - \Pr(Z_{\infty S} \leq \theta)\} - \beta\{\Pr(Z_{\infty A} \geq \theta) - \Pr(Z_{GA} \geq \theta)\}.$$

Since $\alpha \geq 0$, $\beta \geq 0$, it suffices to show

$$\Pr(Z_{GS} \leq \theta) > \Pr(Z_{\infty S} \leq \theta) \quad \text{and} \quad \Pr(Z_{GA} \leq \theta) < \Pr(Z_{\infty A} \leq \theta) \qquad (\dagger).$$

Bounding $\Pr(Z_{GS} \leq \theta)$ below and $\Pr(Z_{GA} \leq \theta)$ above by Lemma 15, and computing the right hand side of each inequality in (†) explicitly by Lemma 14, it follows that (†) is implied by

$$F_{Y_{GS}}(\theta) \cdot V(0) \; > \; S^+ (V(\theta - 1) - V(\theta)) + V(\theta)$$

and

$$F_{Y_{GA}}(\theta) \cdot (V(\theta) - V(0)) + V(0) \; < \; A^+ (V(\theta - 1) - V(\theta)) + V(\theta).$$

A little arithmetic brings these to the form in the statement of the theorem. ∎

8 Critical Review

In this section we offer a critique of various accounts of catecholamine effects, including our own. First, we review the problems with accounts that are based upon the signal-to-noise ratio (SNR). But we fill in a bit more detail, notably the relation between the mean firing rate and the average power. We also explain the sense in which such accounts are correct. Second, we point out inadequacies in our own argument. These fall into two classes: unsupported assumptions, and areas requiring additional insights.

8.1 Problems with SNR-Based Accounts

Near the end of Section 4.1, we criticized SNR-based accounts of catecholamine effects. Then at the end of Section 4.3, we suggested that they had some redeeming features. We now proceed to review and unify these arguments.

First, let us establish the connection between signal power, firing rates, and the SNR. The SNR arises in electrical engineering theory [19] when considering the extraction of some continuous-time signal $s(t)$ from a noisy background $n(t)$. In an effort to quantify how difficult a task this presents to a receiver, we compare the average power input to the receiver in the presence of signal, $\mathcal{S} = \langle (s(t)+n(t))^2 \rangle$, with the average power in the absence of signal, $\mathcal{N} = \langle n(t)^2 \rangle$. Here the angle brackets represent time-averaging, and the quantities being averaged are the squares of the incident amplitudes, since these are proportional to the incident energy. If $s(t)$ is just a tiny perturbation added on to $n(t)$, then $\mathcal{S} \approx \mathcal{N}$, and the signal will be difficult to detect. On the other hand, if the signal amplitude is high and the noise amplitude is low, then $\mathcal{S} \gg \mathcal{N}$. Hence the quotient $\mathcal{S}/\mathcal{N}$, called the *signal-to-noise ratio*, is a measure of the difficulty of the detection task.

These ideas can be related to our work as follows. We are concerned with repeated attempts to detect a given state of the world, or target event. Consider a single-unit network. Recalling our earlier notation, we write x_S and x_A for the nominal input

in the signal present and absent cases. Thus the network output in these cases is $y_S = f_G(x_S)$ or $y_A = f_G(x_A)$ respectively.

Now each of y_S and y_A represents a *firing rate*; the number of spikes or action potentials per unit time. Let us assume that each spike delivers energy U ergs to its afferent neurons. Then the power (energy per unit time) delivered in each case is $U \cdot y_S$ or $U \cdot y_A$ respectively.

But as we discussed before, these firing rates are usually corrupted by noise. Hence over many trials, the network output is a random variable, Y_S or Y_A. Thus the statistic for measuring the distinguishability of signal present and absent cases is the ratio between the average power per trial in these two cases. Writing $\varepsilon[R]$ for the expectation of an rv R, we have

$$\varepsilon[U \cdot Y_S] = U \cdot \varepsilon[Y_S] = U \cdot \mu(Y_S) \quad \text{and} \quad \varepsilon[U \cdot Y_A] = U \cdot \mu(Y_A).$$

Thus the SNR is $\mu(Y_S)/\mu(Y_A)$.

Now we return to the performance of a single-unit receiver. The crux of our criticism is that increased SNR at the output of a single-unit network does not imply improved performance. For it is easy to find rvs X_S and X_A, and an activation family $\{f_G\}$, where increasing the gain simultaneously drives $\mu(Y_S)$ up and $\mu(Y_A)$ down. This surely causes the SNR to rise. But by the Constant Optimal Performance Theorem, the performance at optimal threshold remains the same.

Yet there is a sense in which this analysis is correct. This is best appreciated by understanding how it goes *wrong* in the case of a single-unit receiver. The problem is that the effect of gain increase upon ρ_{Y_S} and ρ_{Y_A} is not captured by the mean alone. Gain changes will in general alter the shapes of these pdfs, possibly driving apart the main concentrations of probability mass, but simultaneously extending their tails for a countervailing effect. The erroneous intuition that separating the means will improve performance arises from the assumption that the effect of a gain increase is to translate the output pdfs rigidly away from one another.

Yet suppose for a minute that this were so. Then we should expect improved perfomance, at any fixed threshold lying between the means. For in general, translating the signal absent pdf down will reduce the chance of a false alarm (unless the portion that is slid across the threshold is identically zero), and likewise for upward translation of the signal present pdf, and the probability of a miss.

Though none of our network models match this situation exactly, there is an extension that comes close. This is the case of the multi-unit chain, which is illustrated in Figure 12. Here W is the output of a multi-unit network. The output noise rv V and final output rv Z are defined exactly as in a single-unit chain.

Now if W is the output of a network that contains a large number of units, the pdfs ρ_{W_S} and ρ_{W_A} lose most of their internal structure, which is derived from X_S and X_A. Each is an exceedingly narrow, sharply peaked Gaussian, centered respectively on $\mu(Y_S)$ and $\mu(Y_A)$. Thus they closely approximate "delta functions" located at these values. Hence the effect of adding output noise V is to create output pdfs ρ_{Z_S} and ρ_{Z_A} that are essentially copies of ρ_V centered at the means. Thus, to a very good

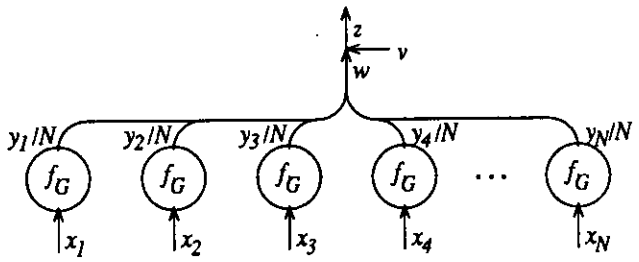

Figure 12: A Multi-Unit Chain

approximation, increasing the gain in this case *does* amount to sliding apart two rigid rvs. This is precisely the intuition behind the chain effect.

8.2 Problems with Our Analysis

Now we critique our own explanation of these effects. These comments are intended to highlight where we have made assumptions because too little is known about the actual operation of the brain, or areas of our investigation where we believe there are additional insights to be garnered.

The single most important assumption concerns our analysis of chains, where we treat the noise rv V as gain independent. One can produce plausible arguments that the variance of this rv should rise, fall or remain constant with increasing gain.

If the variance should fall, we expect that the chain effect would appear even more strongly, since this would presumably reduce the chance of misses and false alarms due to output noise. But if the variance should rise, then these probabilities would presumably increase. Whether this would be offset by the increased separation of the means depends upon the precise behavior of V, W_S and W_A. Thus relaxing the assumption that V is gain-independent will not wholly undermine our argument, but requires us to carry out a more detailed analysis.

Our second comment concerns the ensemble effect. We have argued, on the basis of limited numerical evidence, that the biological import of this effect is small. But this does not constitute a decisive argument, and as we have seen, there are circumstances in which the magnitude of the effect is large. It is an open mathematical problem whether there are biologically plausible cases in which the ensemble effect, or something similar, substantially influences the performance of the network. And it is an interesting neurobiological problem whether such an effect, if present, determines some aspect of behavior.

Whatever the fate of the ensemble effect, there is a flip side to it that is possibly even more interesting. In exactly the same way that a gain increase can be shown in certain circumstances to induce a performance improvement, it is possible to show that in other circumstances an increase will drive down performance. We call this the *anti-ensemble effect*. It is tantalizing to speculate about the countervailing influences of the chain effect and the anti-ensemble effect. For instance, their interaction may

contribute to the fact that human performance improves to a point under the influence of central nervous system stimulants, but then begins to degrade.

9 Summary and Conclusion

In this paper we have proven three theorems about the effect of gain variation upon the signal detection performance of neural networks. The first of these demonstrates that under arbitrary alterations of the activation function, the performance at optimal threshold of a single-unit network is constant. This is the Constant Optimal Performance Theorem. The second states that in spite of this, the optimal performance of a multi-unit network can improve with increasing gain, providing the network is large enough, and the signal detection task is of the proper form. This is the Ensemble Performance Theorem. The third result, the Chain Performance Theorem, states that under suitable assumptions about noise added to the output of one unit, which serves as the input to another, increasing the gain again improves performance. We call the improvement arising from the second theorem the *ensemble effect*, and that arising from the third theorem the *chain effect*.

These results were established under extremely general assumptions about the activation function, and the probability distributions of the input. This is significant because we have based a theory of catecholamine effects upon them [12]. Our claim is not that we have a precise and accurate model of the brain, but rather that our results are sufficiently general and encompassing that whatever model is actually correct, the effects that we have identified explain the way the model's performance varies with increasing gain.

This work makes three major contributions. The first is the demonstration that the influence of catecholamine release upon signal detection performance cannot be understood as a consequence of the effect of these substances upon a single isolated unit. This follows from the Constant Optimal Performance Theorem. The second is the identification of the ensemble effect and the chain effect. These effects arise from interactions among a collection of neurons, assembled either in parallel or in series, and operating in the presence of noise. The Ensemble Performance Theorem and the Chain Performance Theorem explain how a collection of neurons can have signal detection properties that a single neuron lacks. The third contribution is the comparison of the magnitude of these effects.

As a secondary contribution, we have established a framework for further investigation in this area, and shown how to reason within it. Moreover, although the magnitude of the ensemble effect is too small to explain the performance impact of catecholamines, we believe it may yet have a role to play in understanding some aspect of behavior.

Acknowledgements

We wish to thank Jonathan Cohen for his contributions to this work, Allan Heydon and Roni Rosenfeld for unusually thorough readings of early drafts of this report, and Ruth Douglas, John Lehoczky and Halil Mete Soner for guiding us to previous work in large deviation theory.

References

[1] Louis A. Chiodo and Theodore W. Berger. Interactions between dopamine and amino acid-induced excitation and inhibition in the striatum. *Brain Research*, 375:198–203, 1986.

[2] Kai Lai Chung. *A Course in Probability Theory*. Academic Press, second edition, 1974.

[3] Jack R. Cooper, Floyd E. Bloom, and Robert H. Roth. *The Biochemical Basis of Neuropharmacology*. Oxford University Press, fifth edition, 1986.

[4] William Feller. *An Introduction to Probability Theory and Its Applications, Volume II*. Wiley, 1966.

[5] S. L. Foote, R. Freedman, and A. P. Olivier. Effects of putative neurotransmitters on neuronal activity in monkey auditory cortex. *Brain Research*, 86:229–242, 1975.

[6] David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, 1966.

[7] M. E. Munroe. *Introductory Real Analysis*. Addison Wesley, 1965.

[8] L. J. Peloquin and R. Klorman. Effects of methylphenidate on normal children's mood, event-related potentials, and performance in memory scanning and vigilance. *Journal of Abnormal Psychology*, 95:88–98, 1986.

[9] Gordon Raisbeck. *Information Theory: An Introduction for Scientists and Engineers*. The M.I.T. Press, 1964.

[10] H. L. Royden. *Real Analysis*. Macmillan, second edition, 1968.

[11] Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, third edition, 1976.

[12] D. Servan-Schreiber, H. Printz, and J. Cohen. A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. To appear in *Science*, 1990.

[13] David Servan-Schreiber. Information processing models of catecholamine modulation of behavior. June 1989. Thesis Proposal.

[14] I. S. Sokolnikoff and R. M. Redheffer. *Mathematics of Physics and Modern Engineering*. McGraw-Hill, second edition, 1966.

[15] D. W. Stroock. *An Introduction to the Theory of Large Deviations*. Springer-Verlag, 1984.

[16] S. R. S. Varadhan. *Large Deviations and Applications. Regional Conference Series in Applied Mathematics*, Society for Industrial and Applied Mathematics, 1984.

[17] Alan J. Weir. *Lebesgue Integration and Measure*. Cambridge University Press, 1973.

[18] D. J. Woodward, H. C. Moises, B. D. Waterhouse, B. J. Hoffer, and R. Freedman. Modulatory action of norepinehprine in the central nervous system. *Federation Proceedings*, 38:2109–2116, 1979.

[19] R. E. Ziemer and W. H. Tranter. *Principles of Communications: Systems, Modulation and Noise*. Houghton Mifflin, second edition, 1985.