Introduction to Bayesian Methods - 5

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

1. Bayesian classification

Data X, classes C

this likelihood is defined by training data

$$P(C|X) = \frac{P(X|C)}{P(X)} P(C)$$

the prior is also defined by training data

we can use the prior learning to assign a class to new data

$$C_k = \arg\max_{C_k} \frac{P(X|C_k)}{P(X)} P(C_k) = \arg\max_{C_k} P(X|C_k) P(C_k)$$

Consider a vector of N attributes given as Boolean variables
$\mathbf{x} = \{x_i\}$ and classify the data vectors with a single Boolean variable.

The learning procedure must yield:

$$P(y)$$

it is easy to obtain it as an empirical distribution from a histogram of training class data: y is Boolean, the histogram has just two bins, and a hundred examples suffice to determine the empirical distribution to better than 10%.

$$P(\mathbf{x}|y)$$

there is a bigger problem here: the arguments have 2^{N+1} different values, and we must estimate $2(2^N-1)$ parameters ... for instance, with N = 30 there are more than 2 billion parameters!

How can we reduce the huge complexity of learning?

we assume the conditional independence of the x_n's:
naive Bayesian learning

for instance, with just two attributes

$$P(x_1, x_2 | y) = P(x_1 | x_2, y) P(x_2 | y) = P(x_1 | y) P(x_2 | y)$$

conditional independence assumption

with more than 2 attributes

$$P(\mathbf{x} | y) \approx \prod_{k=1}^{N} P(x_k | y)$$

Therefore:

$$P(y_k|\mathbf{x}) = \frac{P(\mathbf{x}|y_k)}{P(\mathbf{x})}P(y_k) = \frac{P(\mathbf{x}|y_k)}{\sum_j P(\mathbf{x}|y_j)P(y_j)}P(y_k)$$

$$\approx \frac{\prod_{n=1}^{N} P(x_n|y_k)}{\sum_j P(y_j)\prod_{n=1}^{N} P(x_n|y_j)}P(y_k)$$

and we assign the class according to the rule (MAP)

$$y = \arg\max_{y_k} \frac{\prod_{n=1}^{N} P(x_n|y_k)}{\sum_j P(y_j)\prod_{n=1}^{N} P(x_n|y_j)}P(y_k)$$

More general discrete inputs

If any of the N x variables has J different values, e if there are K classes, then we must estimate in all $NK(J\text{-}1)$ free parameters with the Naive Bayes Classifier (this includes normalization) (compare this with the $K(J^N\text{-}1)$ parameters needed by a complete classifier)

Short digression: neural networks and their activation functions

The Perceptron (McCulloch and Pitts, 1943)

Bulletin of Mathematical Biology Vol. 52, No. 1/2, pp. 99–115, 1990.
Printed in Great Britain.

0092–8240/90$3.00 + 0.00
Pergamon Press plc
Society for Mathematical Biology

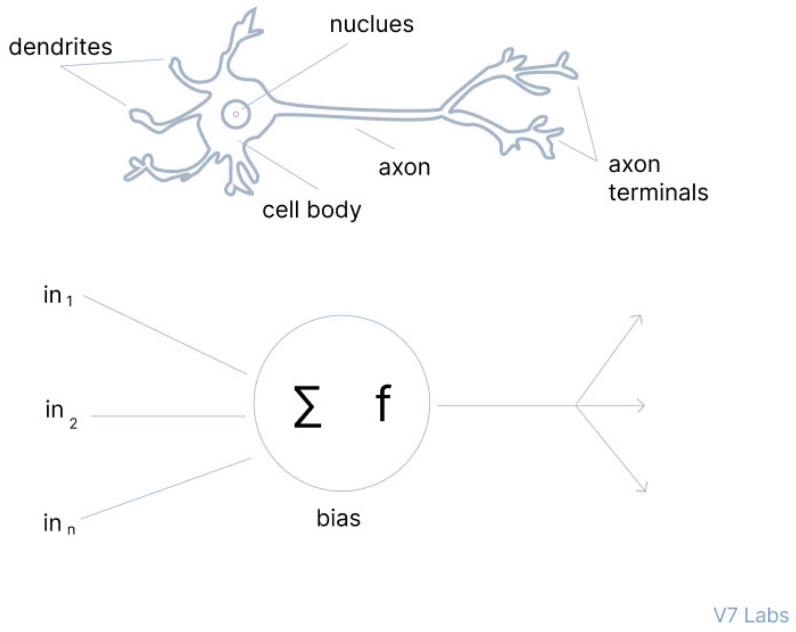
A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

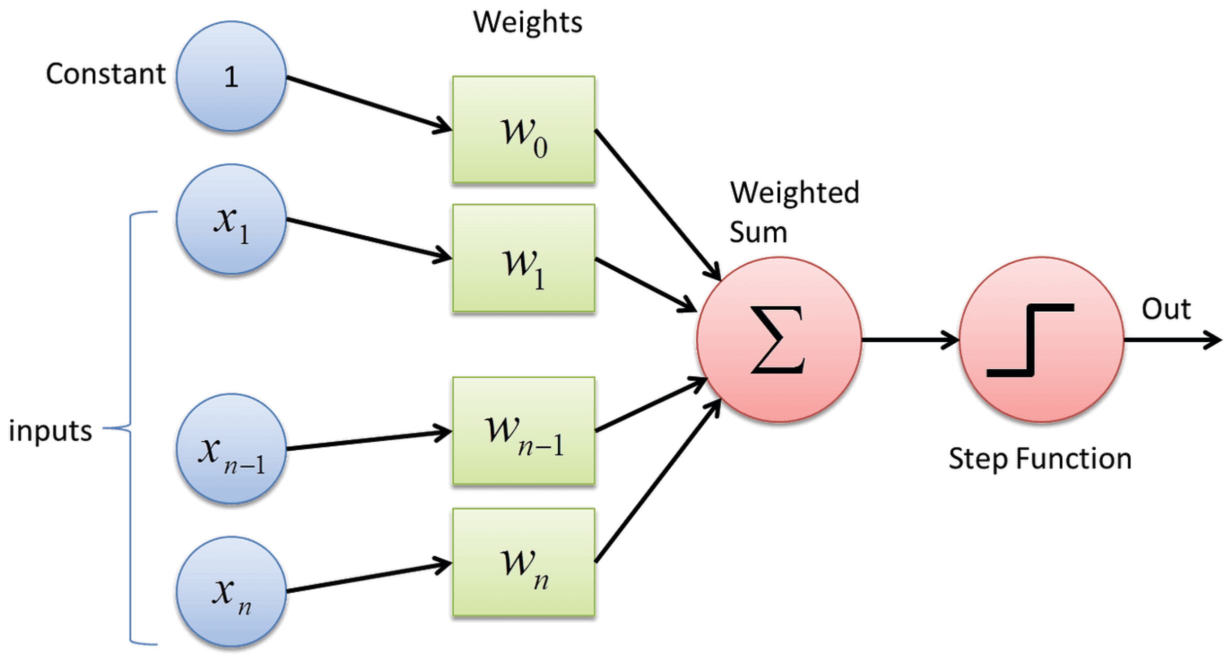
■ WARREN S. MCCULLOCH AND WALTER PITTS
University of Illinois, College of Medicine,
Department of Psychiatry at the Illinois Neuropsychiatric Institute,
University of Chicago, Chicago, U.S.A.

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

Short digression: neural networks and their activation functions

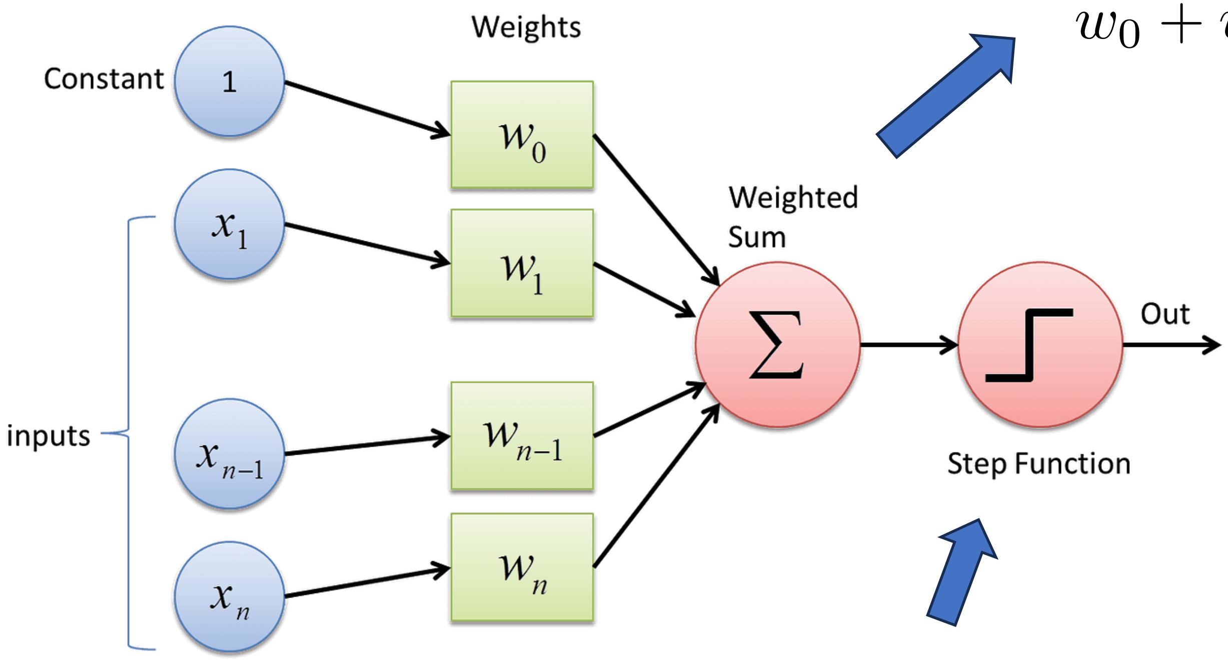
The Perceptron (McCulloch and Pitts, 1943)



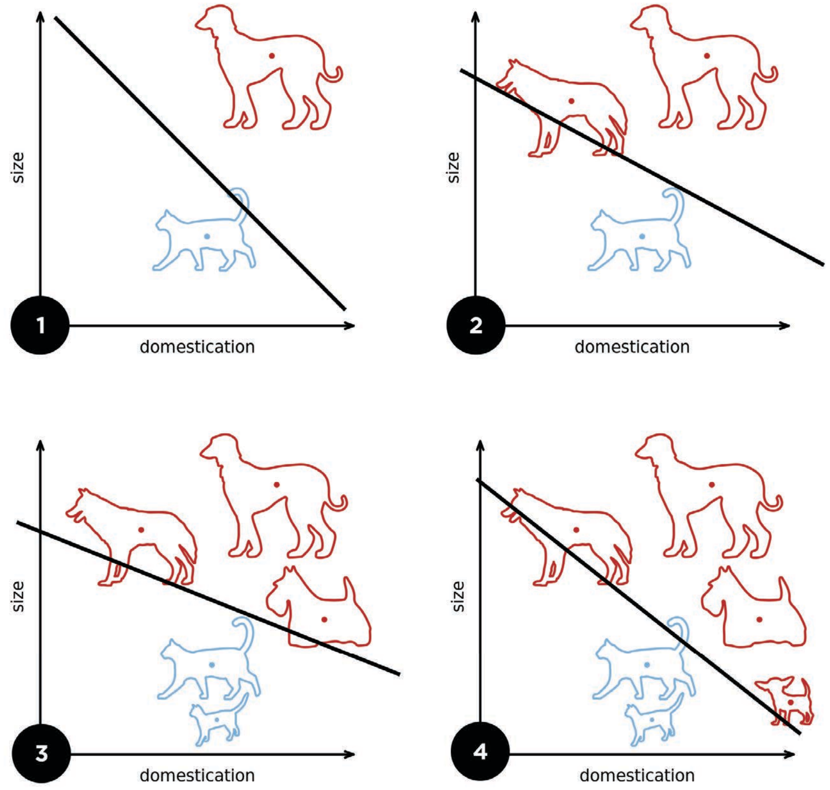


weighted sum

$$w_0 + w_1 x_1 + \cdots + w_{n-1} x_{n-1} + w_n x_n$$



The step function is a specific type of *activation function*



weighted sum = 0 is the equation of a (hyper)plane, and the activation function defines a pair of classes on opposite sides of the (hyper)plane.

"Training" corresponds to selecting the parameters for a correct classification of the examples.

After the training step, the network is fixed and can be used to classify additional inputs.

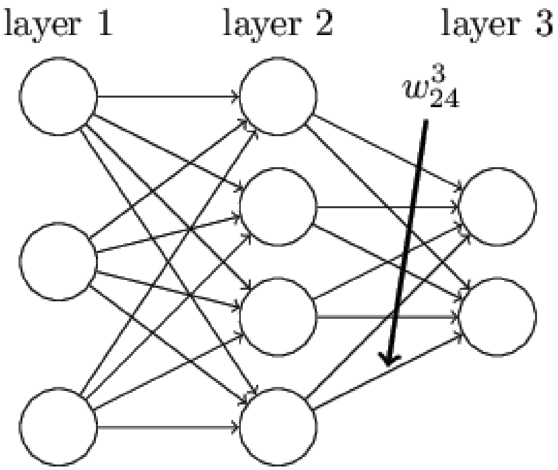
A diagram showing a perceptron updating the plane position as more training examples are added

Frank Rosenblatt
1928–1969

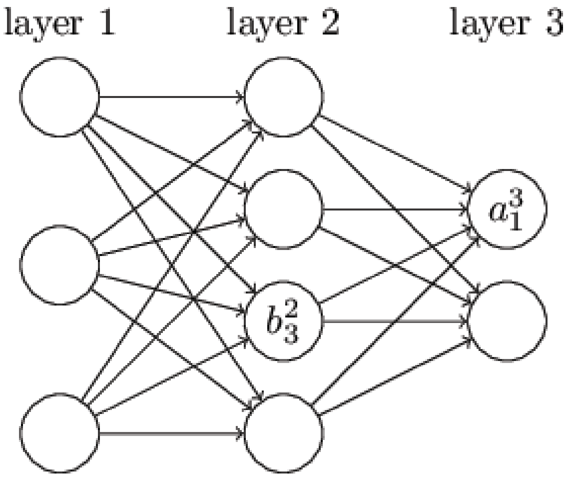
Rosenblatt's perceptron played an important role in the history of machine learning. Initially, Rosenblatt simulated the perceptron on an IBM 704 computer at Cornell in 1957, but by the early 1960s he had built special-purpose hardware that provided a direct, parallel implementation of perceptron learning. Many of his ideas were encapsulated in "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms" published in 1962. Rosenblatt's work was criticized by Marvin Minksy, whose objections were published in the book "Perceptrons", co-authored with Seymour Papert. This book was widely misinterpreted at the time as showing that neural networks were fatally flawed and could only learn solutions for linearly separable problems. In fact, it only proved such limitations in the case of single-layer networks such as the perceptron and merely conjectured (incorrectly) that they applied to more general network models. Unfortunately, however, this book contributed to the substantial decline in research funding for neural computing, a situation that was not reversed until the mid-1980s. Today, there are many hundreds, if not thousands, of applications of neural networks in widespread use, with examples in areas such as handwriting recognition and information retrieval being used routinely by millions of people.

From C. M. Bishop, "Pattern Recognition and Machine Learning",
https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/

Multilayer feedforward networks



w^l_{jk} is the weight from the k^{th} neuron in the $(l-1)^{\text{th}}$ layer to the j^{th} neuron in the l^{th} layer



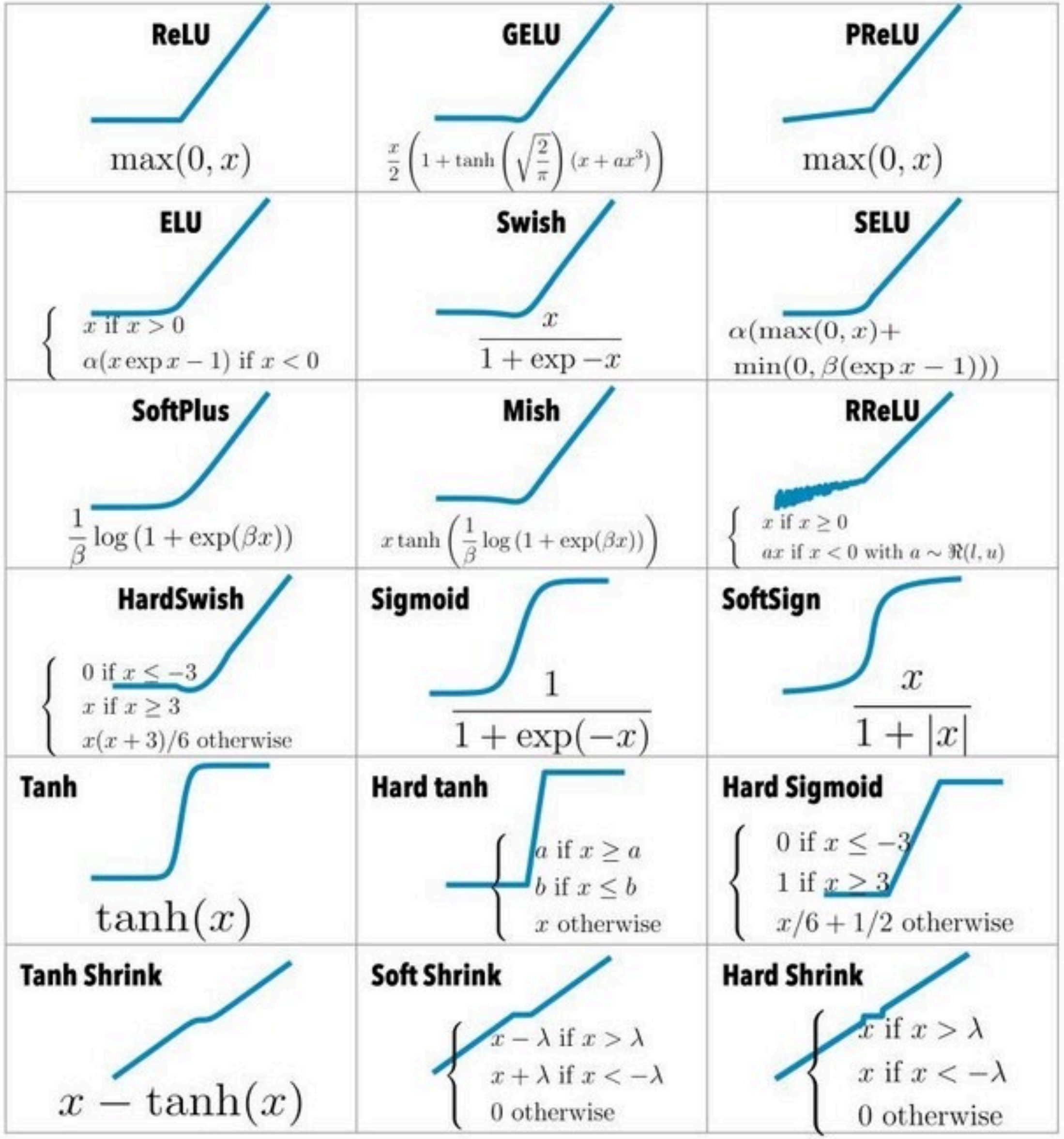
$$a^l_j = \sigma\left(\sum_k w^l_{jk} a^{l-1}_k + b^l_j\right)$$

activation of the j-th neuron in the l-th layer includes the *biases*

σ is the *activation function*

schematics from http://neuralnetworksanddeeplearning.com

Neural Network Activation Functions: a small subset!



Back to Naive Bayesian Learning: Continuous inputs and discrete classes – the Gaussian case

$$P\left(x_n \middle| y_k\right) = \frac{1}{\sqrt{2\pi\sigma_{nk}^2}} \exp\left[-\frac{\left(x_n - \mu_{nk}\right)^2}{2\sigma_{nk}^2}\right]$$

here we must estimate 2*NK* parameters + the shape of the distribution *P*(*y*) (this adds up to another *K*-1 parameters)

Gaussian special case with class-independent variance and Boolean classification (two classes only):

$$P(y=0|\mathbf{x}) = \frac{P(\mathbf{x}|y=0)P(y=0)}{P(\mathbf{x}|y=0)P(y=0)+P(\mathbf{x}|y=1)P(y=1)}$$

$$P(x_n|y=0) = \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left[-\frac{(x_n-\mu_{n0})^2}{2\sigma_n^2}\right]$$

$$P(x_n|y=1) = \frac{1}{\sqrt{2\pi\sigma_n^2}}\exp\left[-\frac{(x_n-\mu_{n1})^2}{2\sigma_n^2}\right]$$

$$P\left(y=0|\mathbf{x}\right)=\frac{P\left(\mathbf{x}|y=0\right)P\left(y=0\right)}{P\left(\mathbf{x}|y=0\right)P\left(y=0\right)+P\left(\mathbf{x}|y=1\right)P\left(y=1\right)}$$

$$=\frac{1}{1+\dfrac{P\left(\mathbf{x}|y=1\right)P\left(y=1\right)}{P\left(\mathbf{x}|y=0\right)P\left(y=0\right)}}$$

$$=\frac{1}{1+\dfrac{P\left(y=1\right)}{P\left(y=0\right)}\prod_{n=1}^{N}\exp\left[-\dfrac{\left(x_n-\mu_{n1}\right)^2}{2\sigma_n^2}+\dfrac{\left(x_n-\mu_{n0}\right)^2}{2\sigma_n^2}\right]}$$

$$=\frac{1}{1+\exp\left\{\ln\left(\dfrac{P\left(y=1\right)}{P\left(y=0\right)}\right)+\sum_{n=1}^{N}\left[\dfrac{\left(\mu_{n1}-\mu_{n0}\right)x_n}{\sigma_n^2}+\dfrac{\mu_{n0}^2-\mu_{n1}^2}{2\sigma_n^2}\right]\right\}}$$

$$w_0 = \ln\left(\frac{P(y=1)}{P(y=0)}\right) + \sum_{n=1}^{N}\left[\frac{\mu_{n0}^2 - \mu_{n1}^2}{2\sigma_n^2}\right]$$

$$w_n = \frac{(\mu_{n1} - \mu_{n0})}{\sigma_n^2}$$

$$\Downarrow$$

logistic shape

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$P(y=1|\mathbf{x}) = 1 - P(y=0|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

Finally, an input vector belongs to class y = 0 if

$$\frac{P(y=0|\mathbf{x})}{P(y=1|\mathbf{x})} > 1$$

$$P(y=0|\mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$P(y=1|\mathbf{x}) = \frac{\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}{1+\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right)}$$

$$\exp\left(w_0 + \sum_{n=1}^{N} w_n x_n\right) < 1$$

$$w_0 + \sum_{n=1}^{N} w_n x_n < 0$$

Back again to neural networks: universality

Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

Abstract. In this paper we demonstrate that finite linear combinations of compositions of a fixed, univariate function and a set of affine functionals can uniformly approximate any continuous function of n real variables with support in the unit hypercube; only mild conditions are imposed on the univariate function. Our results settle an open question about representability in the class of single hidden layer neural networks. In particular, we show that arbitrary decision regions can be arbitrarily well approximated by continuous feedforward neural networks with only a single internal, hidden layer and any continuous sigmoidal nonlinearity. The paper discusses approximation properties of other possible types of nonlinearities that might be implemented by artificial neural networks.

Neural Networks, Vol. 2, pp. 359–366, 1989

0893-6080/89 $3.00 + .00

ORIGINAL CONTRIBUTION

Multilayer Feedforward Networks are Universal Approximators

KUR' HORNIK

Technische Universität Wien

MAXWELL STINCHCOMBE AND HALBEK WHITE

University of California, San Diego

(*Received* 16 *September* 1988; *revised and accepted* 9 *March* 1989)

Abstract—*This paper rigorously establishes that standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. In this sense, multilayer feedforward networks are a class of universal approximators.*

http://neuralnetworksanddeeplearning.com/chap4.html

2. The Li&Ma method

THE ASTROPHYSICAL JOURNAL, **272**:317–324, 1983 September 1

ANALYSIS METHODS FOR RESULTS IN GAMMA-RAY ASTRONOMY

TI-PEI LI AND YU-QIAN MA
High Energy Astrophysics Group, Institute of High Energy Physics,
Academia Sinica, Beijing, China
Received 1982 September 20; accepted 1983 February 7

ABSTRACT

The current procedures for analyzing results of γ-ray astronomy experiments are examined critically. We propose two formulae to estimate the significance of positive observations in searching γ-ray sources or lines. The correctness of the formulae are tested by Monte Carlo simulations.

Subject headings: gamma-rays: general — numerical methods

I. INTRODUCTION

Evaluation of the statistical reliability of positive results in searching discrete γ-ray sources or lines is an important problem in γ-ray astronomy. Since both the signal-to-background ratio and detector sensitivity are generally limited in this energy range, one must carefully analyze the observed data to determine the confidence level of a candidate source or line, that is, the probability that the count rate excess is due to a genuine source or line rather than to a spurious background fluctuation, even though all systematic effects are believed to have been removed.

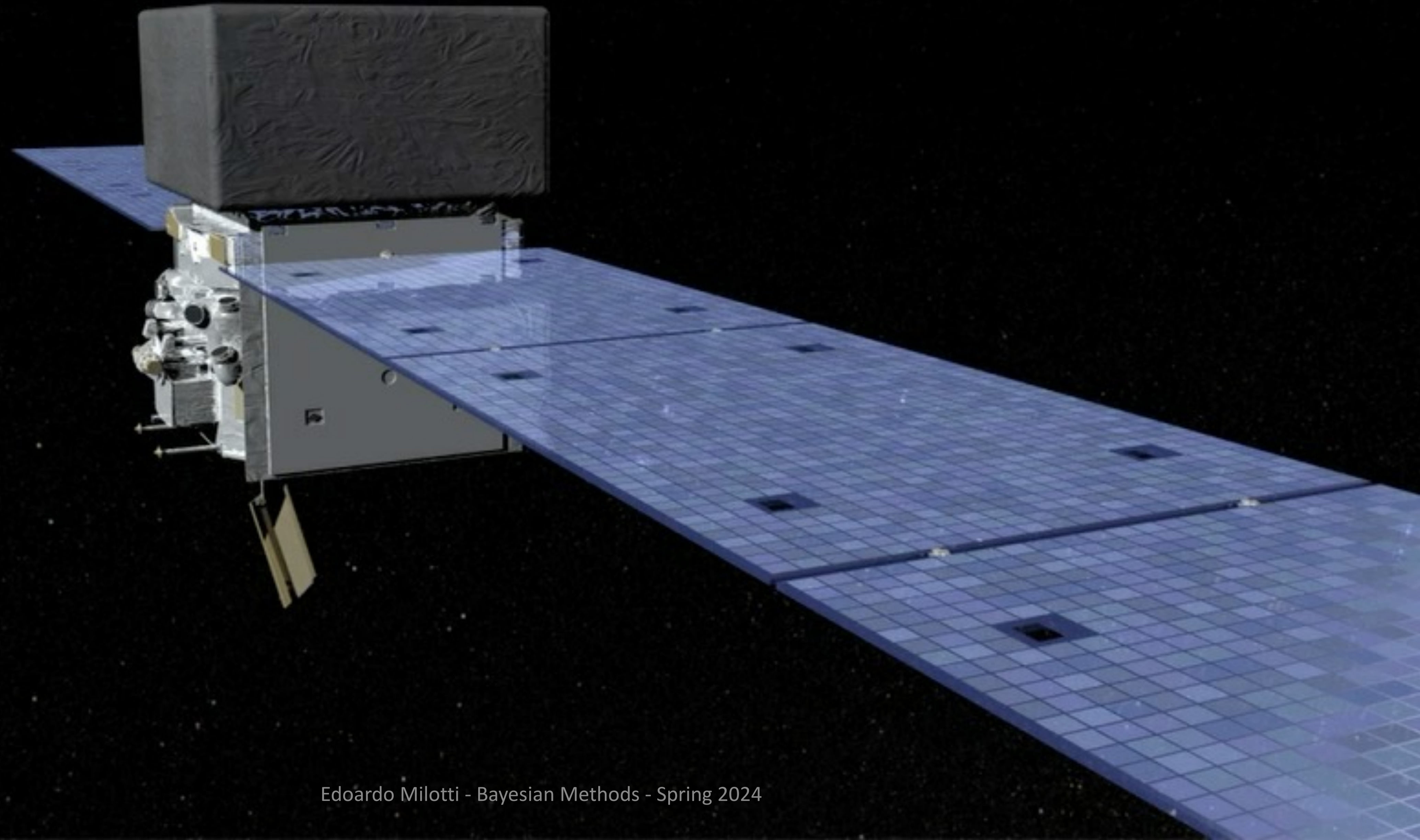
The Fermi Gamma-ray Space Telescope

The Universe is home to numerous exotic and beautiful phenomena, some of which can generate almost inconceivable amounts of energy. Supermassive black holes, merging neutron stars, streams of hot gas moving close to the speed of light ... these are but a few of the marvels that generate gamma-ray radiation, the most energetic form of radiation, billions of times more energetic than the type of light visible to our eyes. What is happening to produce this much energy? What happens to the surrounding environment near these phenomena? How will studying these energetic objects add to our understanding of the very nature of the Universe and how it behaves?

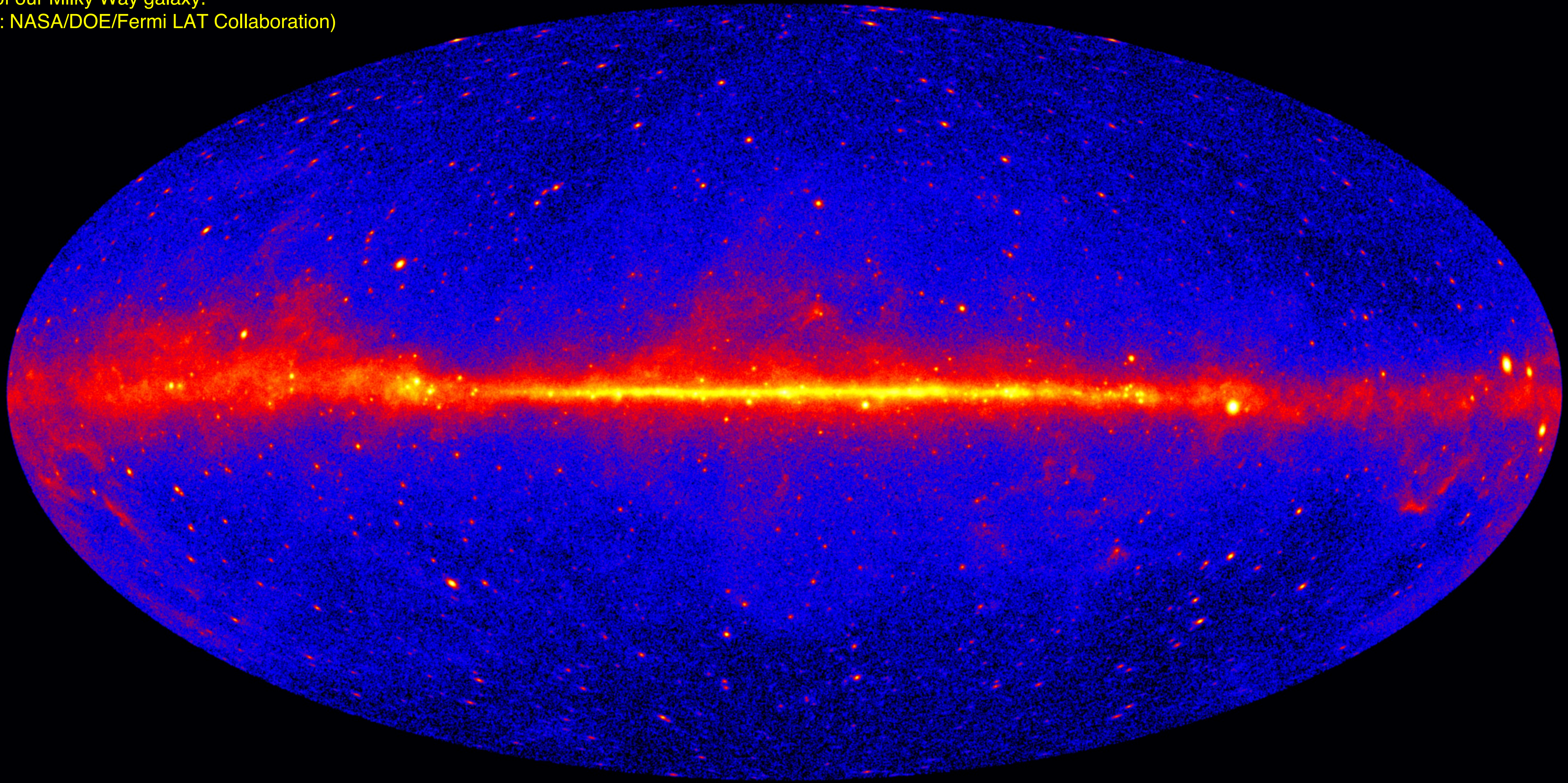
The **Fermi Gamma-ray Space Telescope**, formerly GLAST, is opening this high-energy world to exploration and helping us answer these questions. With Fermi, astronomers have a superior tool to study how black holes, notorious for pulling matter in, can accelerate jets of gas outward at fantastic speeds. Physicists are able to study subatomic particles at energies far greater than those seen in ground-based particle accelerators. And cosmologists are gaining valuable information about the birth and early evolution of the Universe.

(adapted from https://fermi.gsfc.nasa.gov)

https://imagine.gsfc.nasa.gov/observatories/learning/fermi/mission/lat.html



The Fermi LAT 60-month image, constructed from front-converting gamma rays with energies greater than 1 GeV. The most prominent feature is the bright band of diffuse glow along the map's center, which marks the central plane of our Milky Way galaxy.
(Credit: NASA/DOE/Fermi LAT Collaboration)



Gamma-ray (blue) and radio (red) light curves of three millisecond pulsars discovered by radio follow-up in Fermi unidentified sources.

(from https://fermi.gsfc.nasa.gov/science/eteu/pulsars/)

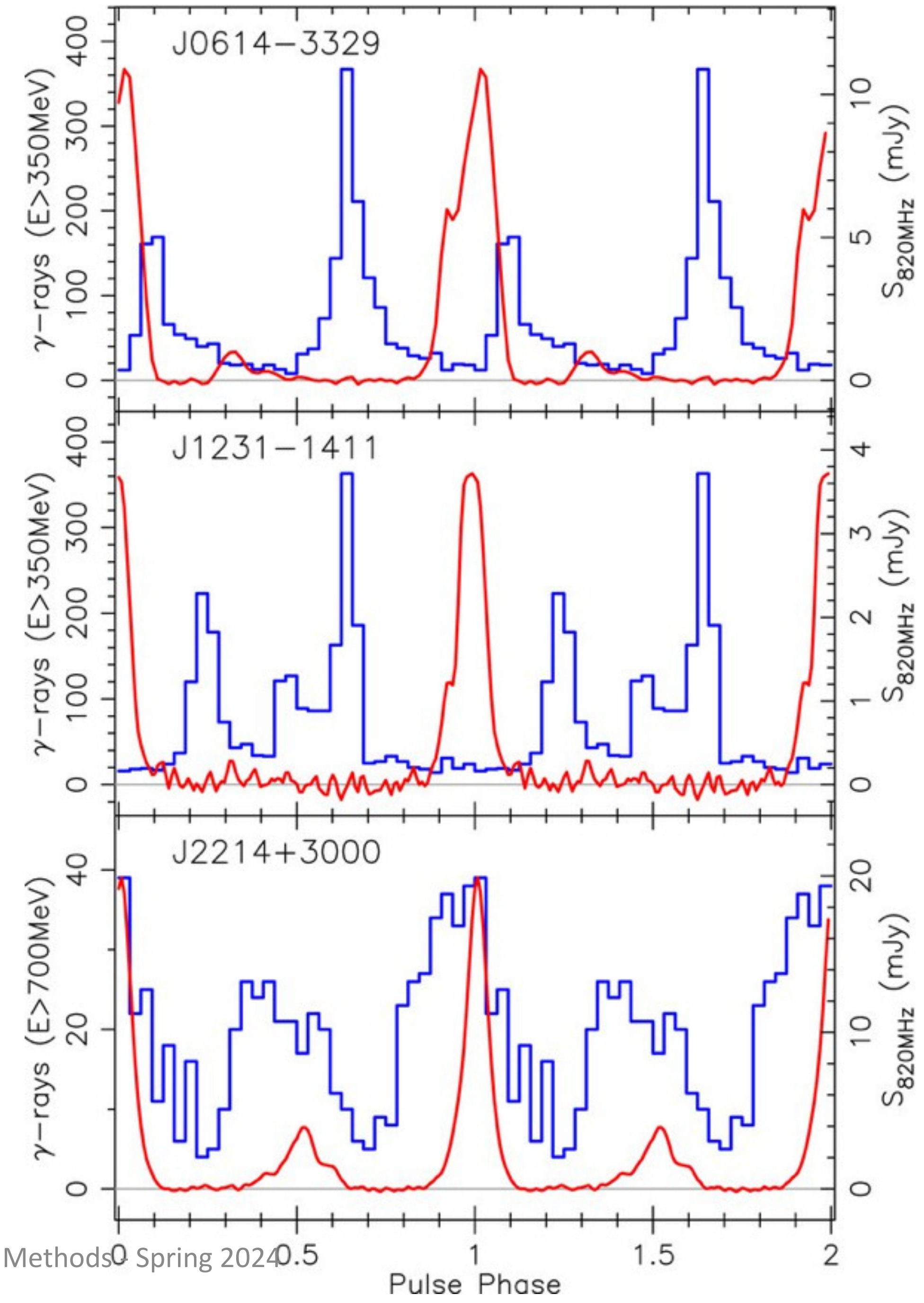


Figure 1 shows a typical observation in γ-ray astronomy. A photon detector points in the direction of a suspected source for a certain time t_{on} and counts N_{on} photons, and then it turns for background measurement for a time interval t_{off} and counts N_{off} photons. The quantity α is the ratio of the on-source time to the off-source time, $\alpha = t_{\mathrm{on}}/t_{\mathrm{off}}$ (in some cases of searching for lines, N_{on} is the number of counts under a peak in an energy spectrum, and the peak is taken to be n_S channels wide; N_{off} is the number of counts in n_b channels adjacent to the peak; then $\alpha = n_S/n_b$).

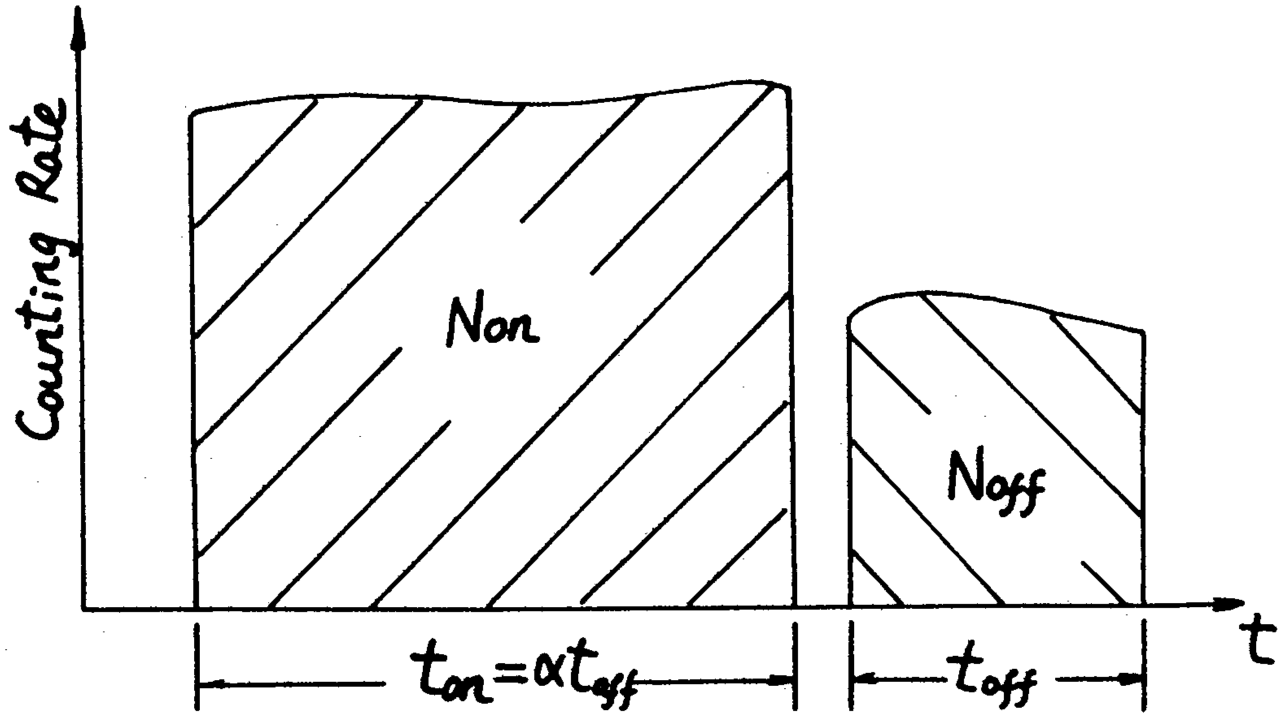


FIG. 1.—A typical observation in γ-ray astronomy

Simple estimate of signal strength and its statistical significance.

- estimate of background photons included in on-source counts

$$\hat{N}_B = \alpha N_{\text{off}}$$

- estimate of observed signal

$$\hat{N}_S = N_{\text{on}} - \hat{N}_B = N_{\text{on}} - \alpha N_{\text{off}}$$

- standard deviation of signal

$$\sigma^2(\hat{N}_S) = \sigma^2(N_{\text{on}}) + \sigma^2(\hat{N}_B) = \sigma^2(N_{\text{on}}) + \sigma^2(\alpha N_{\text{off}}) = \sigma^2(N_{\text{on}}) + \alpha^2\sigma^2(N_{\text{off}})$$

- standard deviation estimate assuming Poisson distr. bkg.

$$\hat{\sigma}_S = \sqrt{N_{\text{on}} + \alpha^2 N_{\text{off}}}$$

- statistical significance

$$S = \frac{\hat{N}_S}{\hat{\sigma}_S} = \frac{N_{\text{on}} - \alpha N \text{off}}{\sqrt{N_{\text{on}} + \alpha^2 N_{\text{off}}}}$$

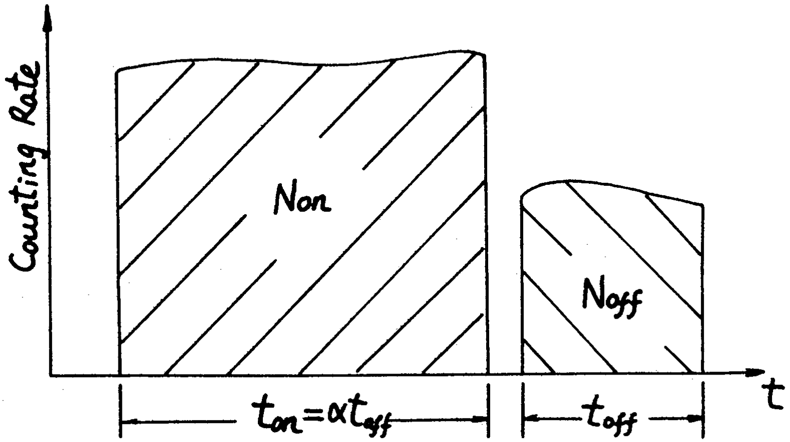


FIG. 1.—A typical observation in γ-ray astronomy

Estimate of result reliability and new estimated significance

Here we calculate the standard deviation under the assumption that there are only background photons.

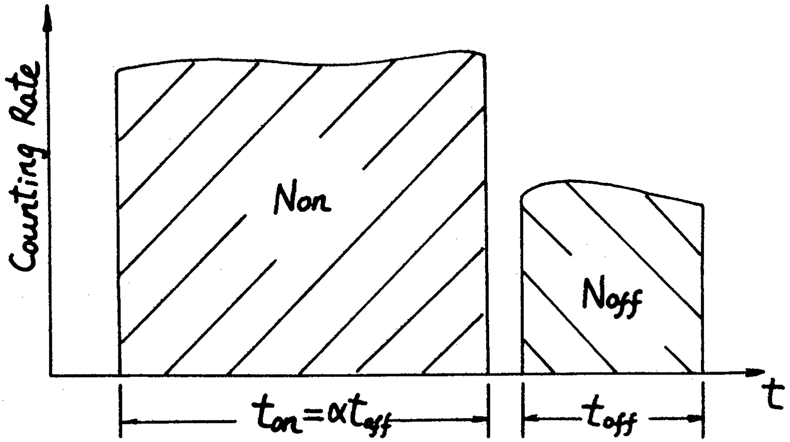


FIG. 1.—A typical observation in γ-ray astronomy

- estimate of photon arrival rate

$$\frac{N_{\rm on}+N_{\rm off}}{t_{\rm on}+t_{\rm off}}$$

- estimate of background on-source photons

$$\hat{N}_B = \frac{N_{\rm on}+N_{\rm off}}{t_{\rm on}+t_{\rm off}} t_{\rm on} = \frac{\alpha}{\alpha+1}(N_{\rm on}+N_{\rm off})$$

- estimate of background off-source photons

$$\frac{N_{\rm on}+N_{\rm off}}{t_{\rm on}+t_{\rm off}} t_{\rm off} = \frac{1}{\alpha+1}(N_{\rm on}+N_{\rm off}) = \frac{\hat{N}_B}{\alpha}$$

- estimate of on-source standard deviation

$$\sigma^2(\hat{N}_S) = \sigma^2(N_{\rm on}) + \alpha^2\sigma^2(N_{\rm off}) \approx \hat{N}_B + \alpha^2(\hat{N}_B/\alpha)$$
$$= (1+\alpha)\hat{N}_B = \alpha(N_{\rm on}+N_{\rm off})$$

- new estimated significance

$$S = \frac{\hat{N}_S}{\hat{\sigma}_S} = \frac{N_{\rm on}-\alpha N_{\rm off}}{(\sqrt{\alpha(N_{\rm on}+N_{\rm off})}}$$

Short recap of the Likelihood Ratio Method (Wilks' theorem) – 1

- Taylor expansion about the MaxL estimator

$$\frac{\partial \ln L(D|\theta)}{\partial \theta} \approx - \left. \frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} (\hat{\theta} - \theta) \approx -E \left[\left. \frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \right] (\hat{\theta} - \theta)$$

- Integration

$$L(D|\theta) \propto \exp \left\{ \frac{1}{2} E \left[\left. \frac{\partial^2 \ln L(D|\theta)}{\partial \theta^2} \right|_{\theta=\hat{\theta}} \right] (\hat{\theta} - \theta)^2 \right\}$$

- Extension to more than one parameters (split into two subsets, recall also the definition of Fisher's information matrix)

$$L(D|\boldsymbol{\theta}) = L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \propto \exp \left[-\frac{1}{2} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T I (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \right]$$

where Fisher's information matrix is split into submatrices $$I = \begin{pmatrix} I_{rr} & \vdots & I_{rs} \\ \cdots & & \cdots \\ I_{sr} & \vdots & I_{ss} \end{pmatrix}$$

Short recap of the Likelihood Ratio Method (Wilks' theorem) – 2

- Then, $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_r \\ \boldsymbol{\theta}_s \end{pmatrix}$ and therefore

$$L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \propto \exp\left[-\frac{1}{2}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)^T I_{rr}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r) - (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)^T I_{rs}(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s) - \frac{1}{2}(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s)^T I_{ss}(\hat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s)\right]$$

- We know that asymptotically, the estimator $\hat{\boldsymbol{\theta}}$ has a Gaussian distribution with covariance matrix I^{-1}, therefore, asymptotically, the likelihood approaches the pdf of the estimator.

- When we maximize the likelihood with respect to the whole parameter vector, we find that the estimators for the subvectors are

$$\theta_r' = \hat{\boldsymbol{\theta}}_r; \quad \theta_s' = \hat{\boldsymbol{\theta}}_s$$

  and the corresponding maximum likelihood has a fixed value that depends only on data.

- When we maximize the likelihood with respect to the *s* parameters only, we find $\theta_s'' = \hat{\boldsymbol{\theta}}_s$ and

$$L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}_s'') \propto \exp\left[-\frac{1}{2}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)^T I_{rr}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)\right]$$

Short recap of the Likelihood Ratio Method (Wilks' theorem) – 3

- This means that when we define the likelihood ratio $\lambda = \dfrac{L(D|\boldsymbol{\theta}_r, \boldsymbol{\theta}''_s)}{L(D|\boldsymbol{\theta}'_r, \boldsymbol{\theta}'_s)}$, and recall that the estimators are

  asymptotically Gaussian, we find that

$$-2 \ln \lambda = (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)^T I_{rr} (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r)$$

  has a chi-square distribution with *r* degrees of freedom (Wilks' theorem).

Application of the Likelihood Ratio Method to estimating N_S and N_B

- The problem at hand is defined by

data: $(N_{\mathrm{on}}, N_{\mathrm{off}})$

unknown parameters: $\boldsymbol{\theta} = (\langle N_B \rangle, \langle N_S \rangle)$

null hypothesis: $\langle N_S \rangle = 0$

alternative hypothesis: $\langle N_S \rangle \neq 0$

- maximum of a Poisson likelihood with just one measurement (*N*)

$$L(N|\theta) = \frac{\theta^N}{N!} e^{-\theta} \quad \Rightarrow \quad \ln L(N|\theta) \sim N \ln \theta - \theta \quad \Rightarrow \quad \frac{\partial L}{\partial \theta} = \frac{N}{\theta} - 1 = 0 \quad \Rightarrow \quad \hat{\theta} = N$$

(the actual measurement is the MaxL estimate).
This means that the previous estimates ARE MaxL estimates, and we can use them to calculate the likelihood ratio.

Application of the Likelihood Ratio Method to estimating N_S and N_B (ctd.)

- **MaxL estimates**

alternative hypothesis: $$\langle \hat{N}_B \rangle = \alpha N_{\mathrm{off}}, \quad \langle \hat{N}_S \rangle = N_{\mathrm{on}} - \alpha N_{\mathrm{off}}$$

null hypothesis: $$\langle \hat{N}_B \rangle = \frac{\alpha}{\alpha+1}(N_{\mathrm{on}} + N_{\mathrm{off}}), \quad \langle \hat{N}_S \rangle = 0$$

- **Likelihoods**

alternative hypothesis: $$L(D|H_1)|_{\mathrm{max}} = \frac{N_{\mathrm{on}}^{N_{\mathrm{on}}}}{N_{\mathrm{on}}!} e^{-N_{\mathrm{on}}} \frac{N_{\mathrm{off}}^{N_{\mathrm{off}}}}{N_{\mathrm{off}}!} e^{-N_{\mathrm{off}}}$$

null hypothesis: $$\begin{aligned} L(D|H_0)|_{\mathrm{max}} = {} & \frac{1}{N_{\mathrm{on}}!} \left(\frac{\alpha}{\alpha+1}(N_{\mathrm{on}} + N_{\mathrm{off}}) \right)^{N_{\mathrm{on}}} \exp\left(-\frac{\alpha}{\alpha+1}(N_{\mathrm{on}} + N_{\mathrm{off}}) \right) \\ & \times \frac{1}{N_{\mathrm{off}}!} \left(\frac{1}{\alpha+1}(N_{\mathrm{on}} + N_{\mathrm{off}}) \right)^{N_{\mathrm{off}}} \exp\left(-\frac{1}{\alpha+1}(N_{\mathrm{on}} + N_{\mathrm{off}}) \right) \end{aligned}$$

Application of the Likelihood Ratio Method to estimating N_S and N_B (ctd.)

- **MaxL ratio**

$$\lambda_{\max} = \frac{L(D|H_0)|_{\max}}{L(D|H_1)|_{\max}} = \left(\frac{\alpha}{\alpha+1} \frac{N_{\text{on}} + N_{\text{off}}}{N_{\text{on}}} \right)^{N_{\text{on}}} \left(\frac{1}{\alpha+1} \frac{N_{\text{on}} + N_{\text{off}}}{N_{\text{off}}} \right)^{N_{\text{off}}}$$

therefore the significance can be obtained from $-2\ln\lambda_{\max}$ because $-2\ln\lambda$ has a chi-square distribution with 1 degree of freedom (only one parameter – the background rate – matters in the case of null hypothesis, while the alternative hypothesis has two parameters – background rate and source rate) .

- if $x^2 \sim \chi^2(1)$ then $|x| \sim \chi(1)$, and we estimate the significance as

$$S \approx \sqrt{-2\ln\lambda_{\max}} = \sqrt{2} \left\{ N_{\text{on}} \ln\left[\frac{\alpha+1}{\alpha} \left(\frac{N_{\text{on}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] + N_{\text{off}} \ln\left[(\alpha+1) \left(\frac{N_{\text{off}}}{N_{\text{on}} + N_{\text{off}}} \right) \right] \right\}$$

(a perfect match with exp. data gives a vanishing chi, the actual value of chi is an estimate of the size of the fluctuation in terms of standard deviations).

- to be continued ...