Introduction to Bayesian Methods - 5

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

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

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

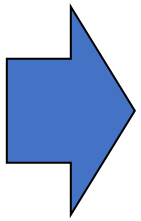
Consider a vector of *N* attributes given as Boolean variables $\mathbf{x} = \{x_i\}$ and classify the data vectors with a single Boolean variable.

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 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)

Continuous inputs and discrete classes – the Gaussian case

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

here we must estimate $2NK$ 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):

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

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

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

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

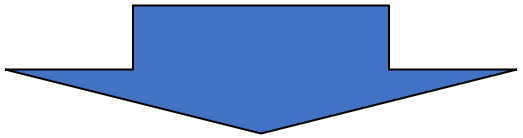
$$= \frac{1}{1+\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)}}$$

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

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

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

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



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

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

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

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

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

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

Bayesian thinking and the human brain

Brit. J. Phil. Sci. **63** (2012), 697–723

Bayes in the Brain—On Bayesian Modelling in Neuroscience

Matteo Colombo and Peggy Seriès

ABSTRACT

According to a growing trend in theoretical neuroscience, the human perceptual system is akin to a Bayesian machine. The aim of this article is to clearly articulate the claims that perception can be considered Bayesian inference and that the brain can be considered a Bayesian machine, some of the epistemological challenges to these claims; and some of the implications of these claims. We address two questions: (i) How are Bayesian models used in theoretical neuroscience? (ii) From the use of Bayesian models in theoretical neuroscience, have we learned or can we hope to learn that perception is Bayesian inference or that the brain is a Bayesian machine? From actual practice in theoretical neuroscience, we argue for three claims. First, currently Bayesian models do not provide mechanistic explanations; instead they are useful devices for predicting and systematizing observational statements about people's performances in a variety of perceptual tasks. That is, currently we should have an instrumentalist attitude towards Bayesian models in neuroscience. Second, the inference typically drawn from Bayesian behavioural performance in a variety of perceptual tasks to underlying Bayesian mechanisms should be understood within the three-level framework laid out by David Marr ([1982]). Third, we can hope to learn that perception *is* Bayesian inference or that the brain *is* a Bayesian machine to the extent that Bayesian models will prove successful in yielding secure and informative predictions of both subjects' perceptual performance and features of the underlying neural mechanisms.

CellPress
REVIEWS

Review

How Do Expectations Shape Perception?

Floris P. de Lange,[1,3,*] Micha Heilbron,[1,3] and Peter Kok[2,3]

Perception and perceptual decision-making are strongly facilitated by prior knowledge about the probabilistic structure of the world. While the computational benefits of using prior expectation in perception are clear, there are myriad ways in which this computation can be realized. We review here recent advances in our understanding of the neural sources and targets of expectations in perception. Furthermore, we discuss Bayesian theories of perception that prescribe how an agent should integrate prior knowledge and sensory information, and investigate how current and future empirical data can inform and constrain computational frameworks that implement such probabilistic integration in perception.

Highlights

Expectations play a strong role in determining the way we perceive the world.

Prior expectations can originate from multiple sources of information, and correspondingly have different neural sources, depending on where in the brain the relevant prior knowledge is stored.

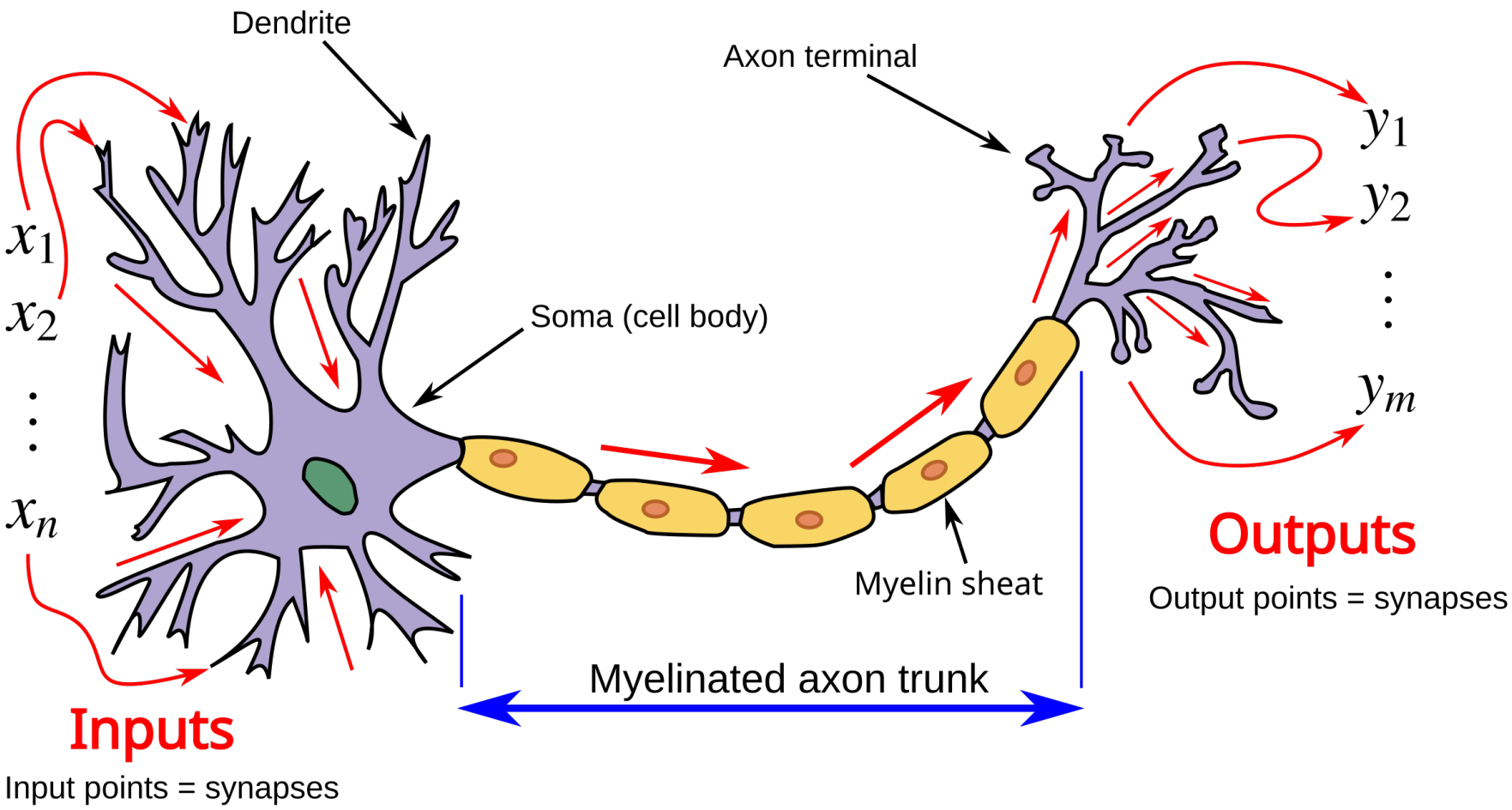
Recent findings from both human neuroimaging and animal electrophysiology have revealed that prior expectations can modulate sensory processing at both early and late stages, and both before and after stimulus onset. The response modulation can take the form of either dampening the sensory representation or enhancing it via a process of sharpening.

Theoretical computational frameworks of neural sensory processing aim to explain how the probabilistic integration of prior expectations and sensory inputs results in perception.

 https://doi.org/10.1016/j.tics.2018.06.002

An *extremely quick* intro to neural networks

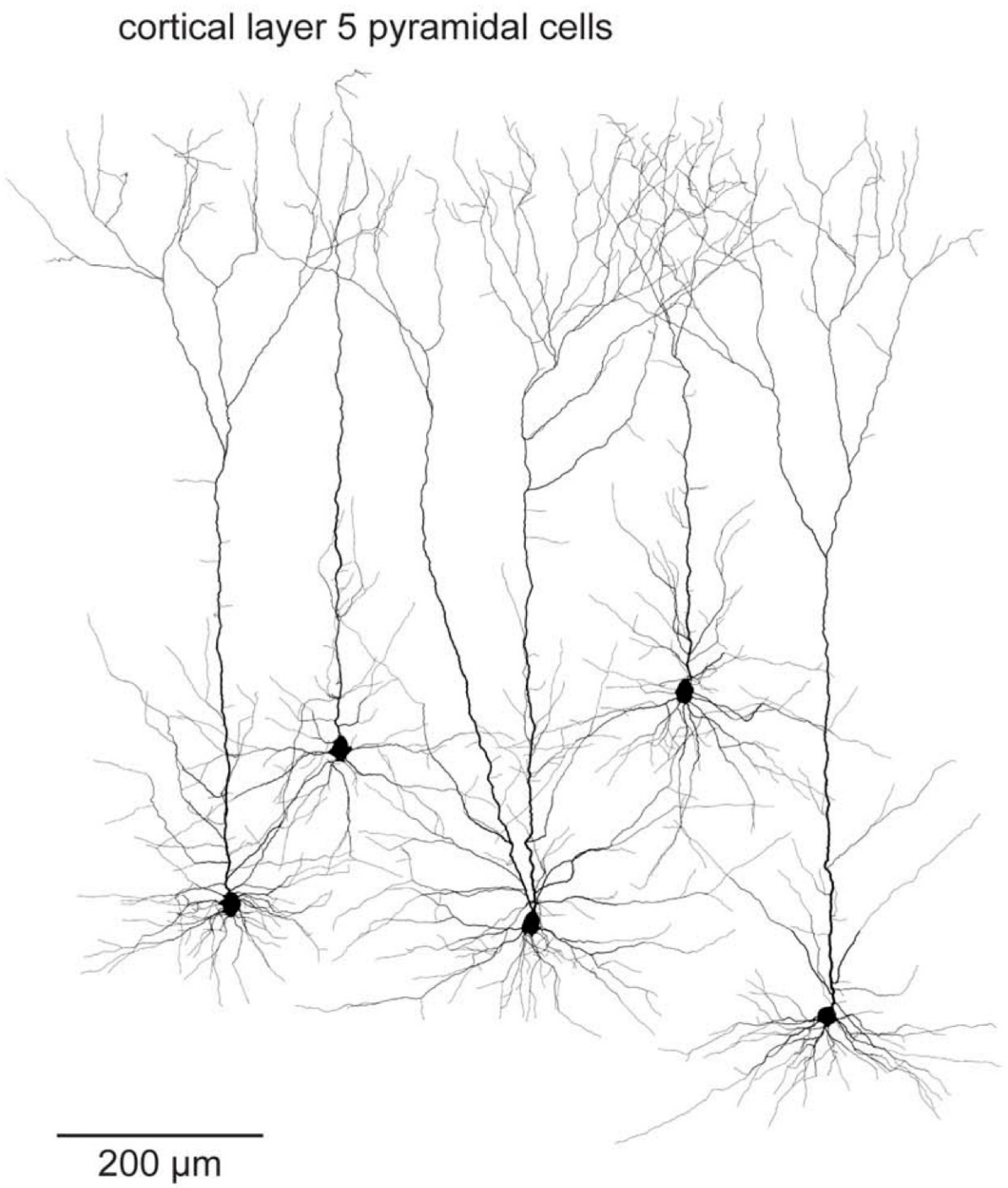
Neurons



Forests of simulated neurons

The synapses of simulated pyramidal cells are indistiguishable from their real counterparts when they are built with the assumption that their morphology minimizes both conduction times within cells and the total cytoplasm mass.

(from Cuntz et al., 10.1371/journal.pcbi.1000877).

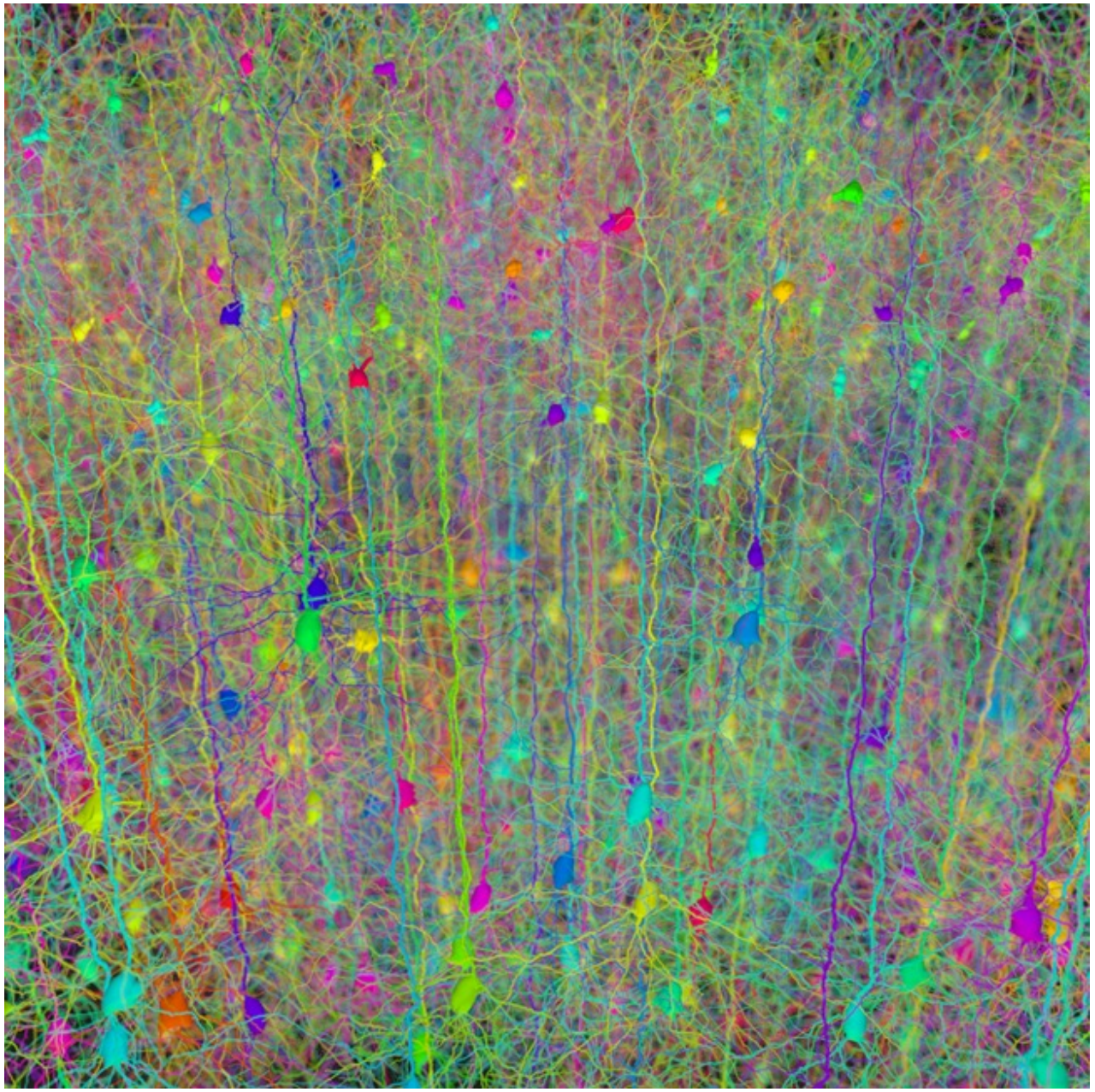


https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000877

Forests of simulated neurons

The synapses of simulated pyramidal cells are indistiguishable from their real counterparts when they are built with the assumption that their morphology minimizes both conduction times within cells and the total cytoplasm mass.

(from Cuntz et al., 10.1371/journal.pcbi.1000877).



PLoS Comput Biol 6(8): ev06.i08. https://doi.org/10.1371/image.pcbi.v06.i08

Towards an understanding of the inner human brain mechanisms

BULLETIN OF
MATHEMATICAL BIOPHYSICS
VOLUME 5, 1943

A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY

WARREN S. MCCULLOCH AND WALTER PITTS

FROM THE UNIVERSITY OF ILLINOIS, COLLEGE OF MEDICINE,
DEPARTMENT OF PSYCHIATRY AT THE ILLINOIS NEUROPSYCHIATRIC INSTITUTE,
AND THE UNIVERSITY OF CHICAGO

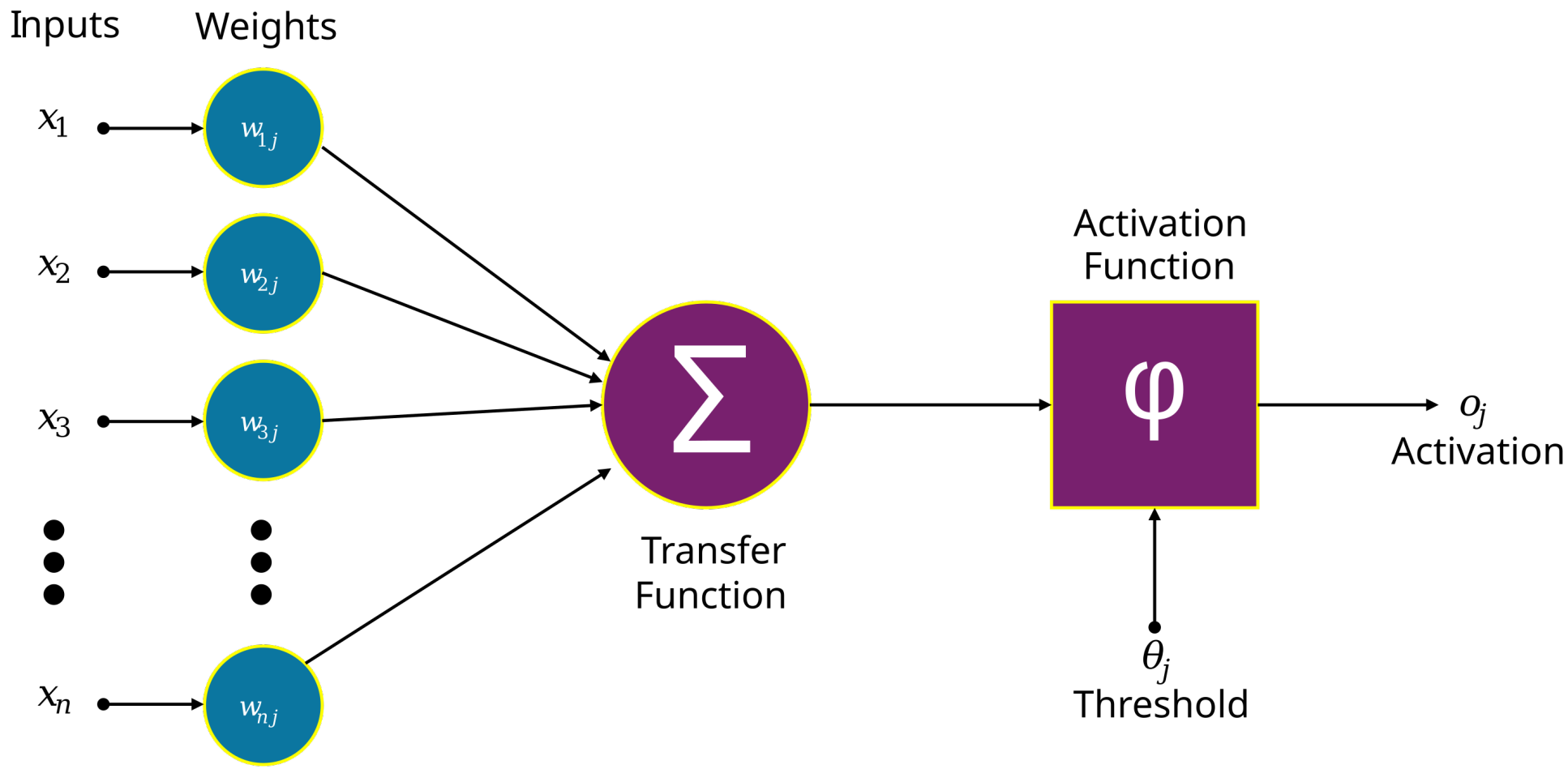
Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; 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.

Carnap, R. 1938. *The Logical Syntax of Language.* New York: Harcourt, Brace and Company.

Hilbert, D., und Ackermann, W. 1927. *Grundüge der Theoretischen Logik.* Berlin: J. Springer.

Russell, B., and Whitehead, A. N. 1925. *Principa Mathematica.* Cambridge: Cambridge University Press.

Synthetic neurons



Logical functions

(the notation follows "Language II", as defined by Rudolf Carnap)

Figure 1a $N_2(t) \;.\equiv.\; N_1(t-1)$

Figure 1b $N_3(t) \;.\equiv.\; N_1(t-1) \;\mathbf{v}\; N_2(t-1)$

Figure 1c $N_3(t) \;.\equiv.\; N_1(t-1).N_2(t-1)$

Figure 1d $N_3(t) \;.\equiv.\; N_1(t-1).\sim N_2(t-1)$

Figure 1e $N_3(t) \;:\equiv:\; N_1(t-1) \;.\mathbf{v}.\; N_2(t-3).\sim N_2(t-2)$

$N_4(t) \;.\equiv.\; N_2(t-2).N_2(t-1)$

Figure 1f $N_4(t) \;:\equiv:\; \sim N_1(t-1).\, N_2(t-1) \;\mathbf{v}\; N_3(t-1) \;.\mathbf{v}.\; N_1(t-1).\, N_2(t-1).\, N_3(t-1)$

$N_4(t) \;:\equiv:\; \sim N_1(t-2).\, N_2(t-2) \;\mathbf{v}\; N_3(t-2) \;.\mathbf{v}.\; N_1(t-2).\, N_2(t-2).\, N_3(t-2)$

Figure 1g $N_3(t) \;.\equiv.\; N_2(t-2).\sim N_1(t-3)$

Figure 1h $N_2(t) \;.\equiv.\; N_1(t-1).N_1(t-2)$

Figure 1i $N_3(t) \;:\equiv:\; N_2(t-1) \;.\mathbf{v}.\; N_1(t-1).(Ex)\,t-1\,.N_1(x)\;.N_2(x)$

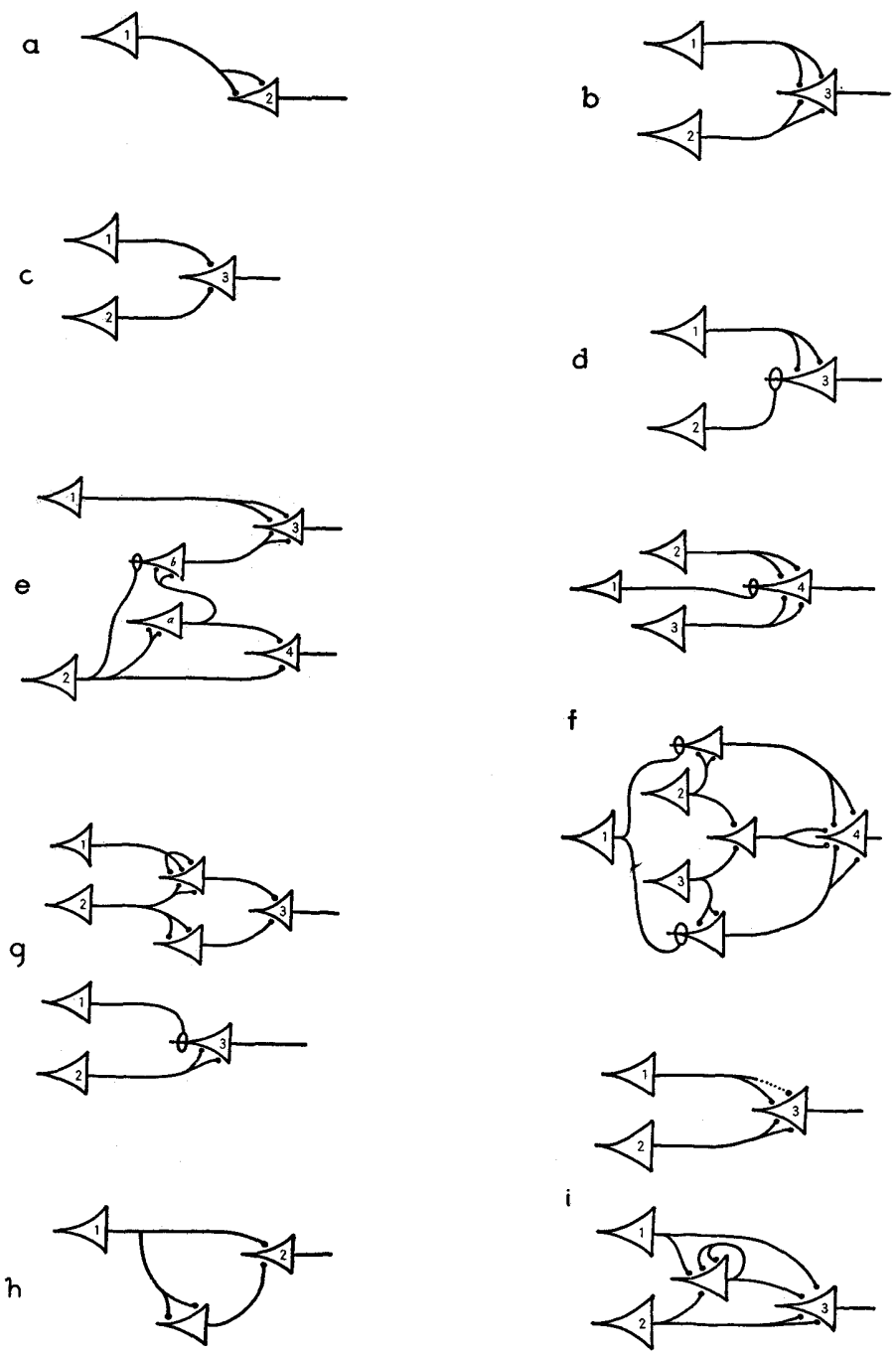
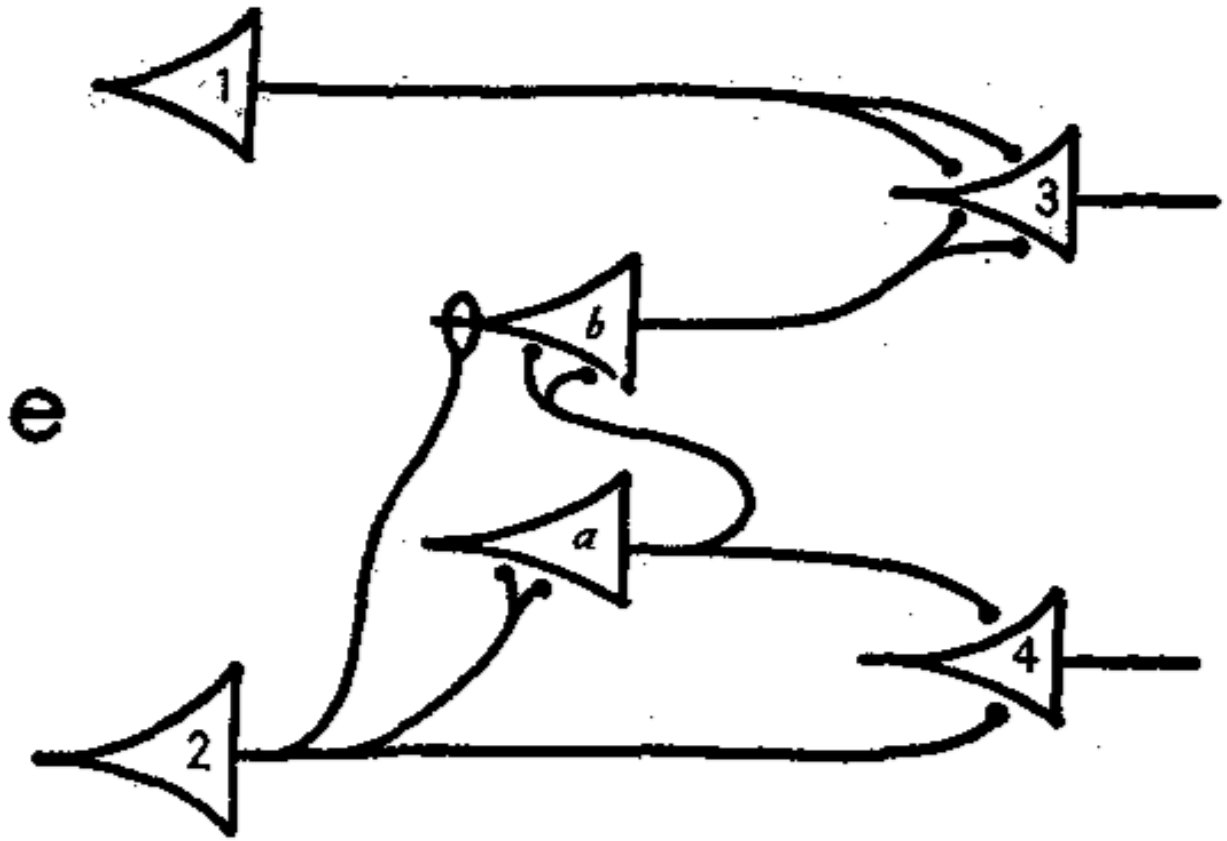


FIGURE 1

Logical functions

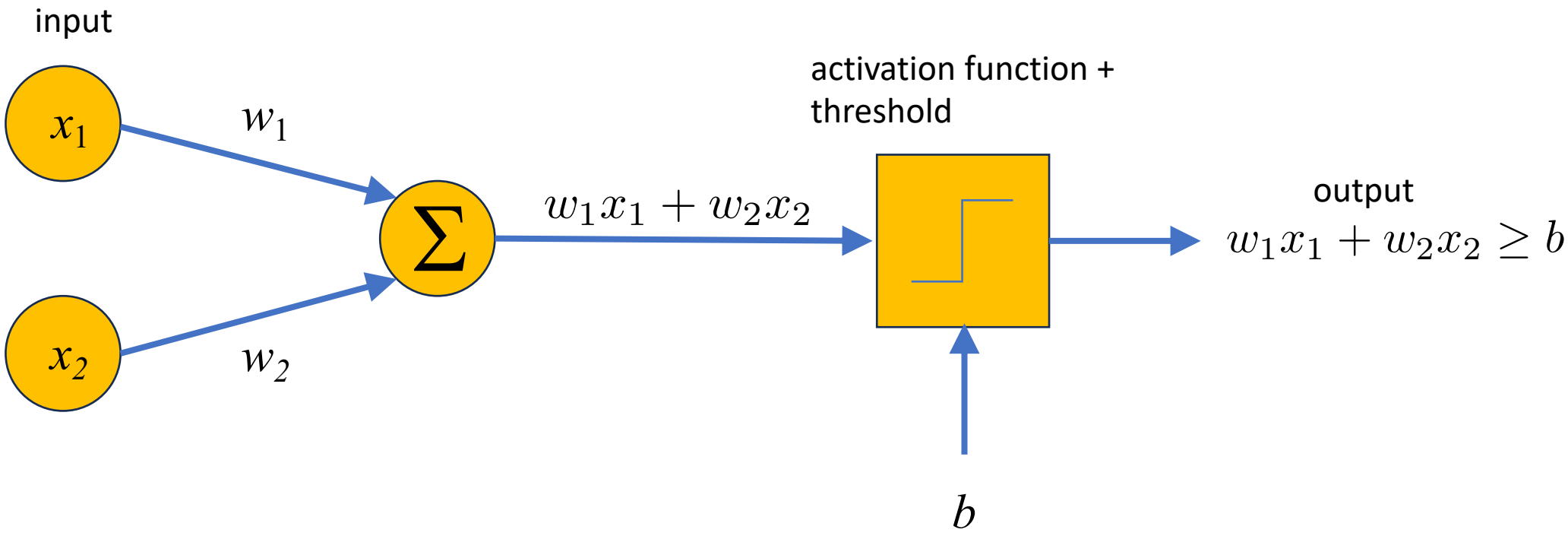
(the notation follows "Language II", as defined by Rudolf Carnap)

$$N_3(t) \ :\equiv : N_1(t-1) \ .\mathbf{v}. \ N_2(t-3). \sim N_2(t-2)$$
$$N_4(t) \ .\equiv . \ N_2(t-2). N_2(t-1)$$



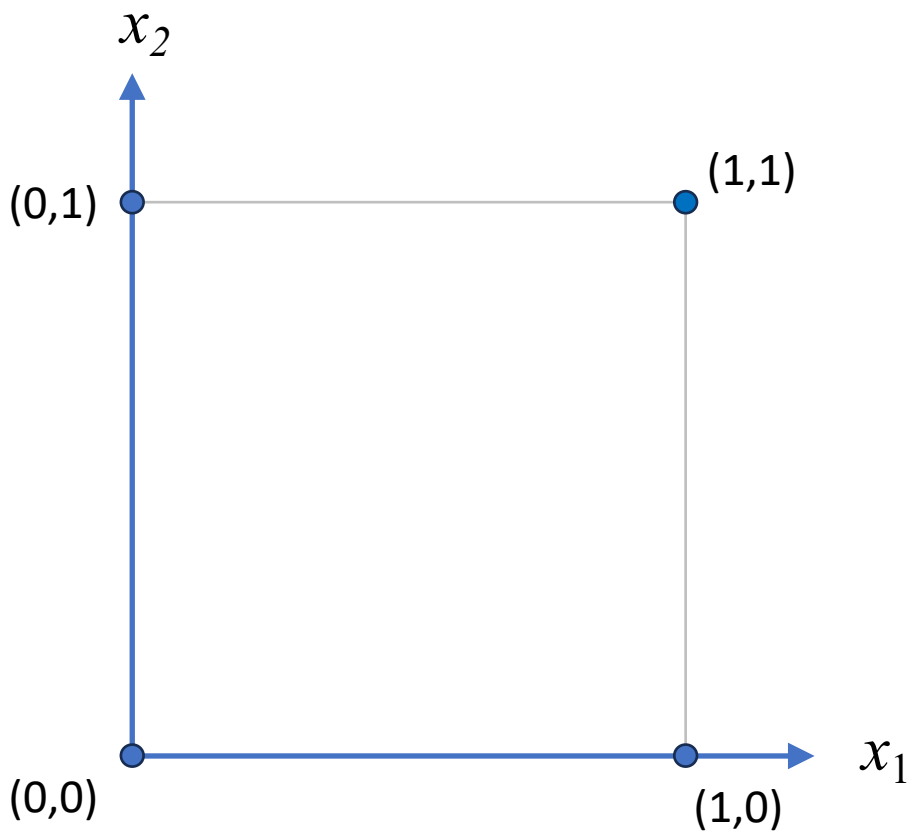
Perceptron

(a simple perceptron with just two inputs)



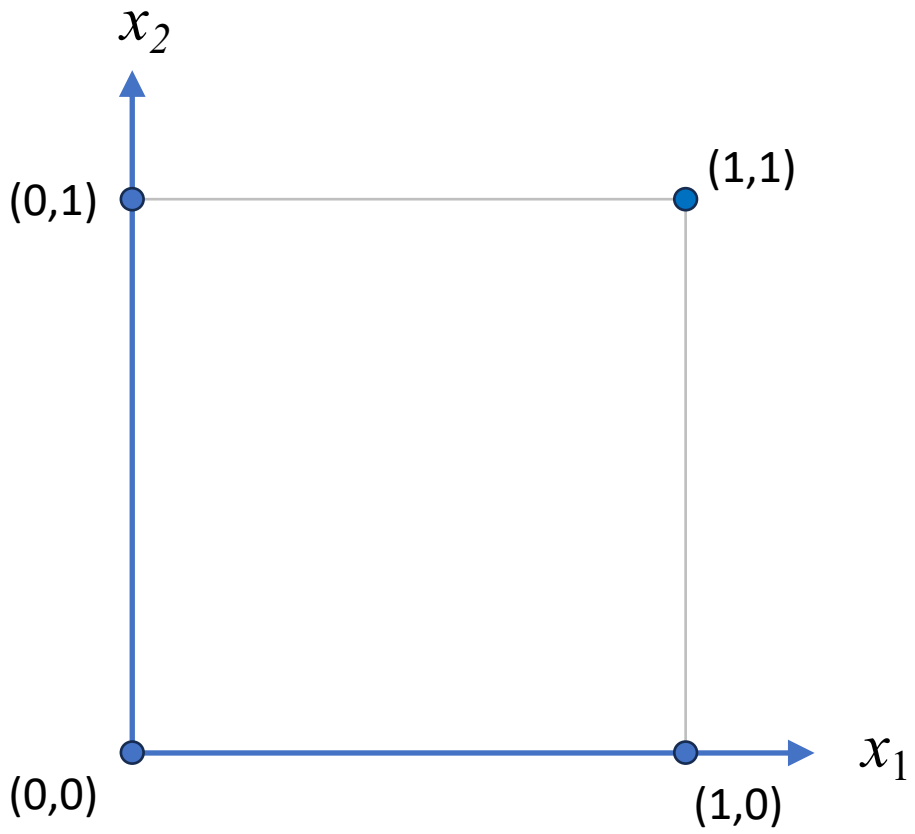
Perceptron

(a simple perceptron with just two inputs)



Perceptron

(a simple perceptron with just two inputs)



If

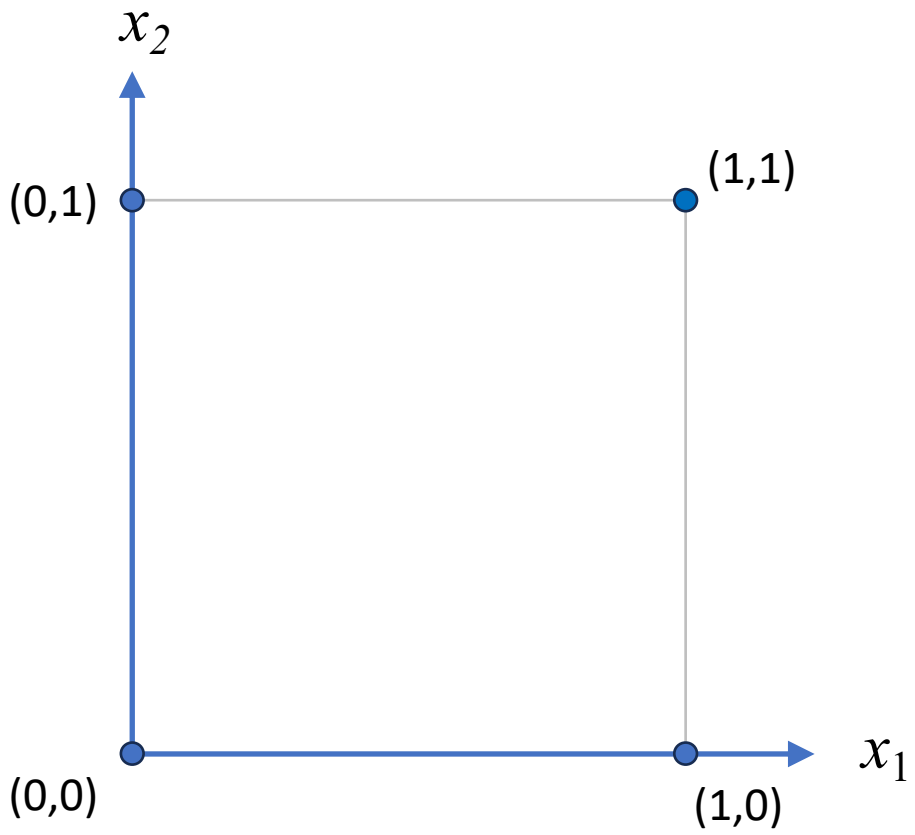
true ⟶ 0

false ⟶ 1

then we find input combinations corresponding to well-defined logical functions.

Perceptron

(a simple perceptron with just two inputs)

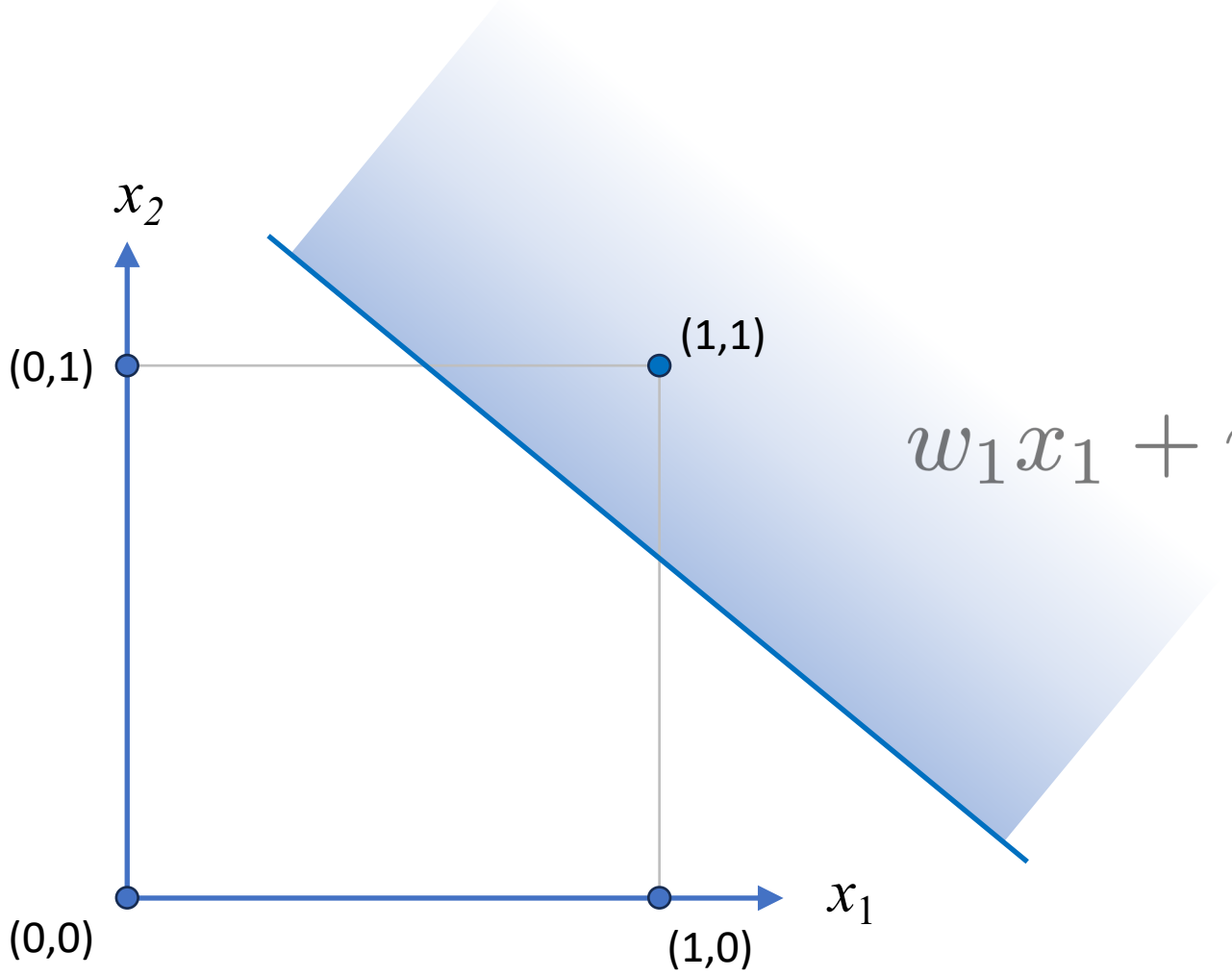


logical AND

x_1	x_2	$x_1 \wedge x_2$
0	0	0
0	1	0
1	0	0
1	1	1

Perceptron

(a simple perceptron with just two inputs)

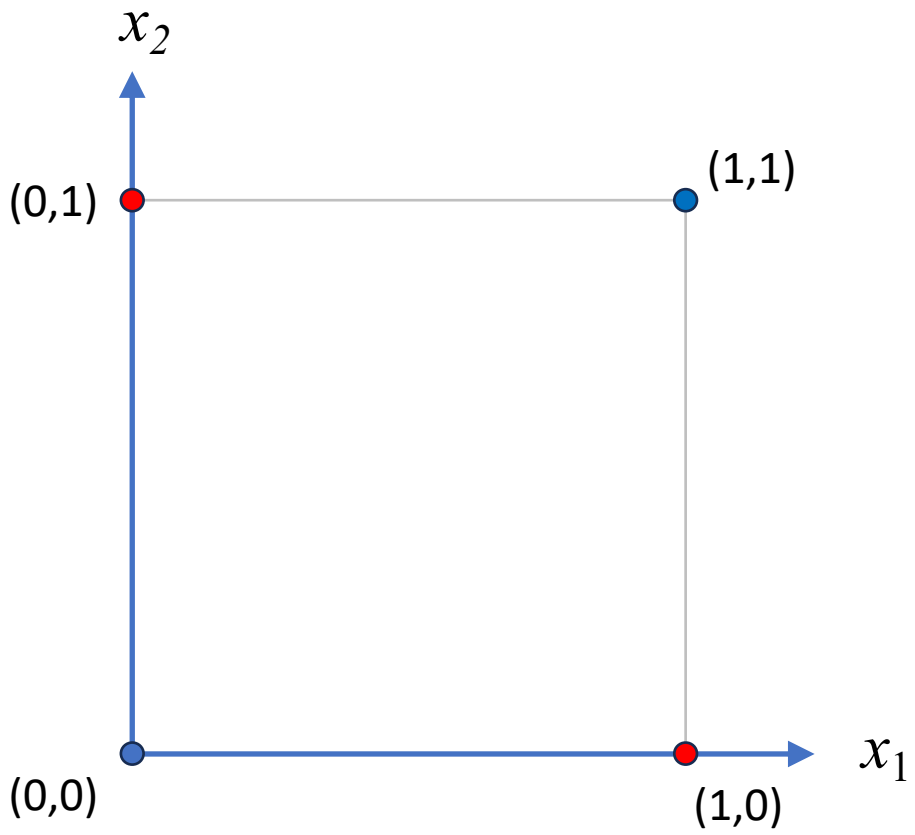


$$w_1 x_1 + w_2 x_2 \geq b$$

$$x_1 \wedge x_2$$

Perceptron

(a simple perceptron with just two inputs)



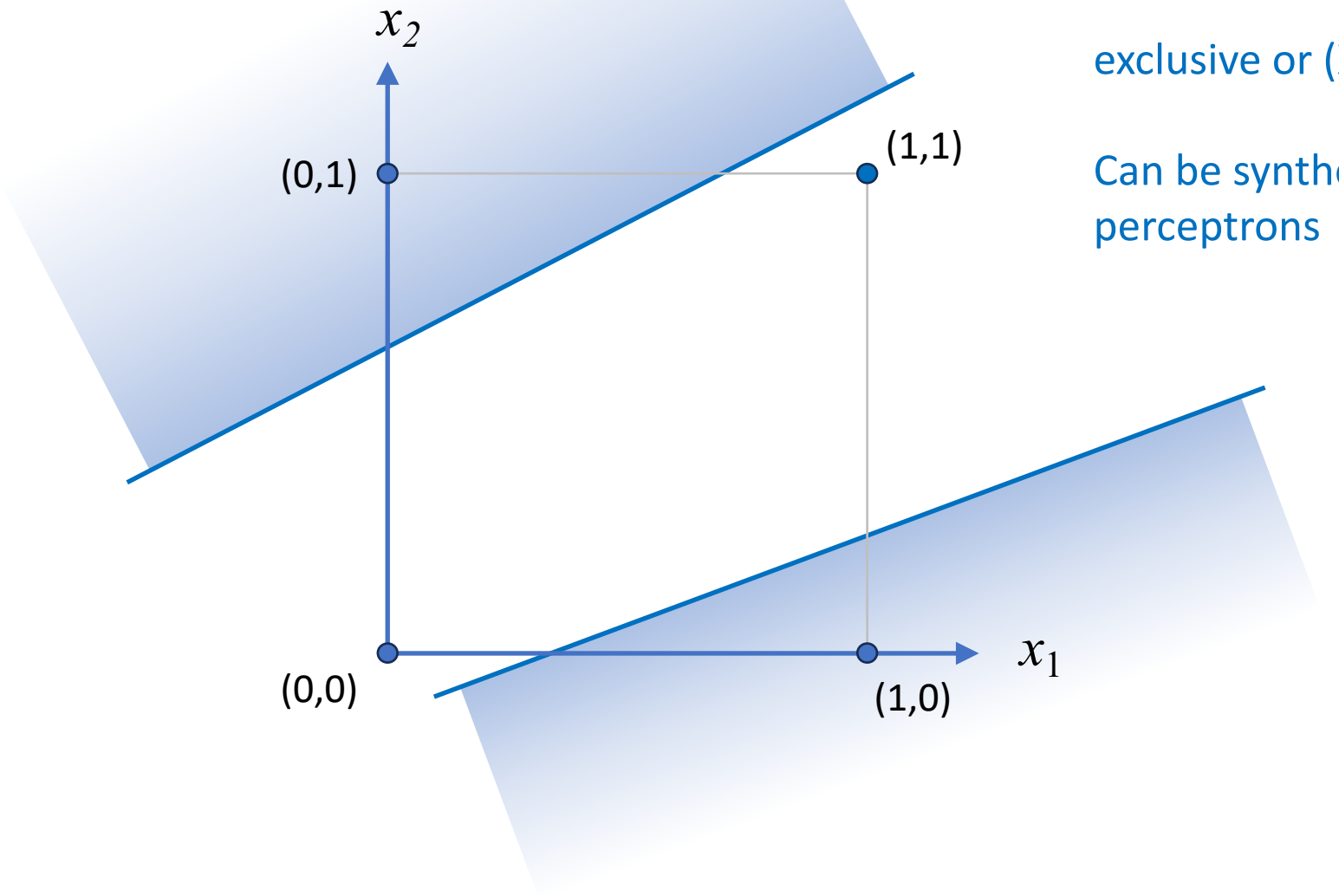
logical XOR

x_1	x_2	x_1 XOR x_2
0	0	0
0	1	1
1	0	1
1	1	0

No simple perceptron corresponds to this logical function!

Two-neuron neural network

This simple network reproduces the XOR function



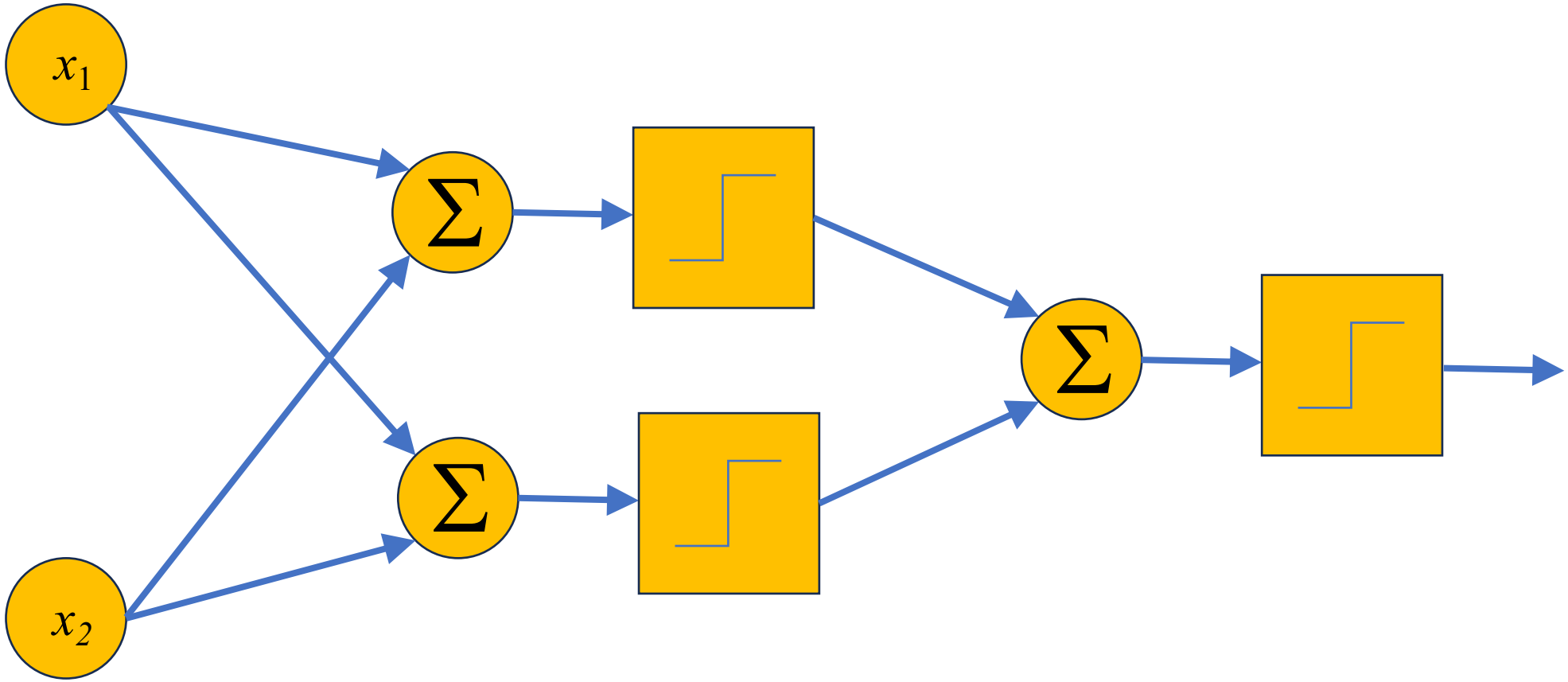
exclusive or (XOR)

Can be synthesized with a pair of perceptrons

Two-neuron neural network

This simple network reproduces the XOR function

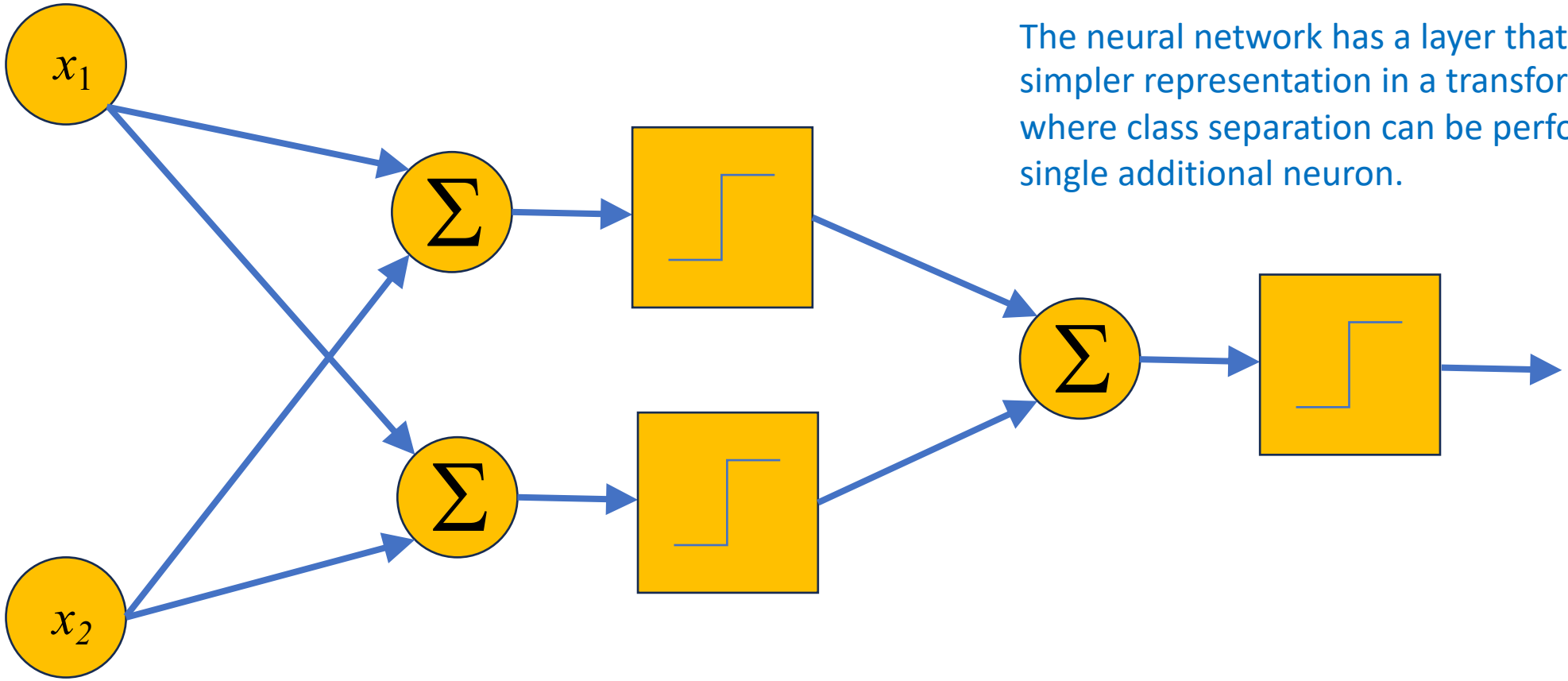
x_1	x_2	$y_1 = x_1 \wedge \bar{x}_2$	$y_2 = \bar{x}_1 \wedge x_2$	$y_1 \vee y_2$
0	0	0	0	0
0	1	0	1	1
1	0	1	0	1
1	1	0	0	0



Two-neuron neural network

This simple network reproduces the XOR function

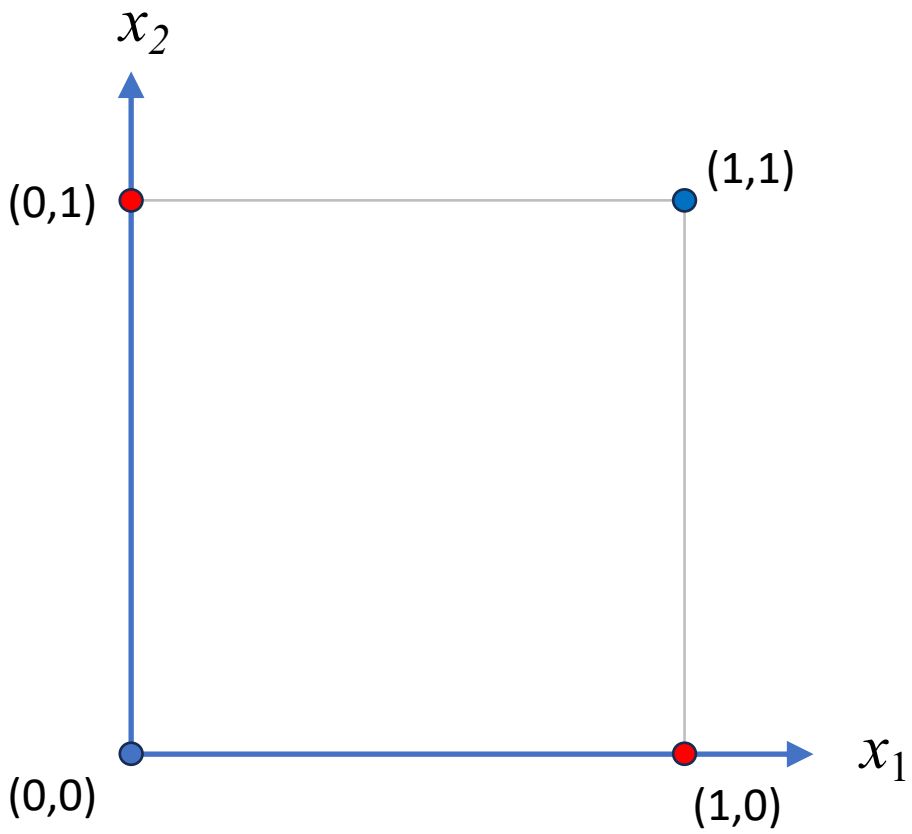
x_1	x_2	$y_1 = x_1 \wedge \bar{x}_2$	$y_2 = \bar{x}_1 \wedge x_2$	$y_1 \vee y_2$
0	0	0	0	0
0	1	0	1	1
1	0	1	0	1
1	1	0	0	0

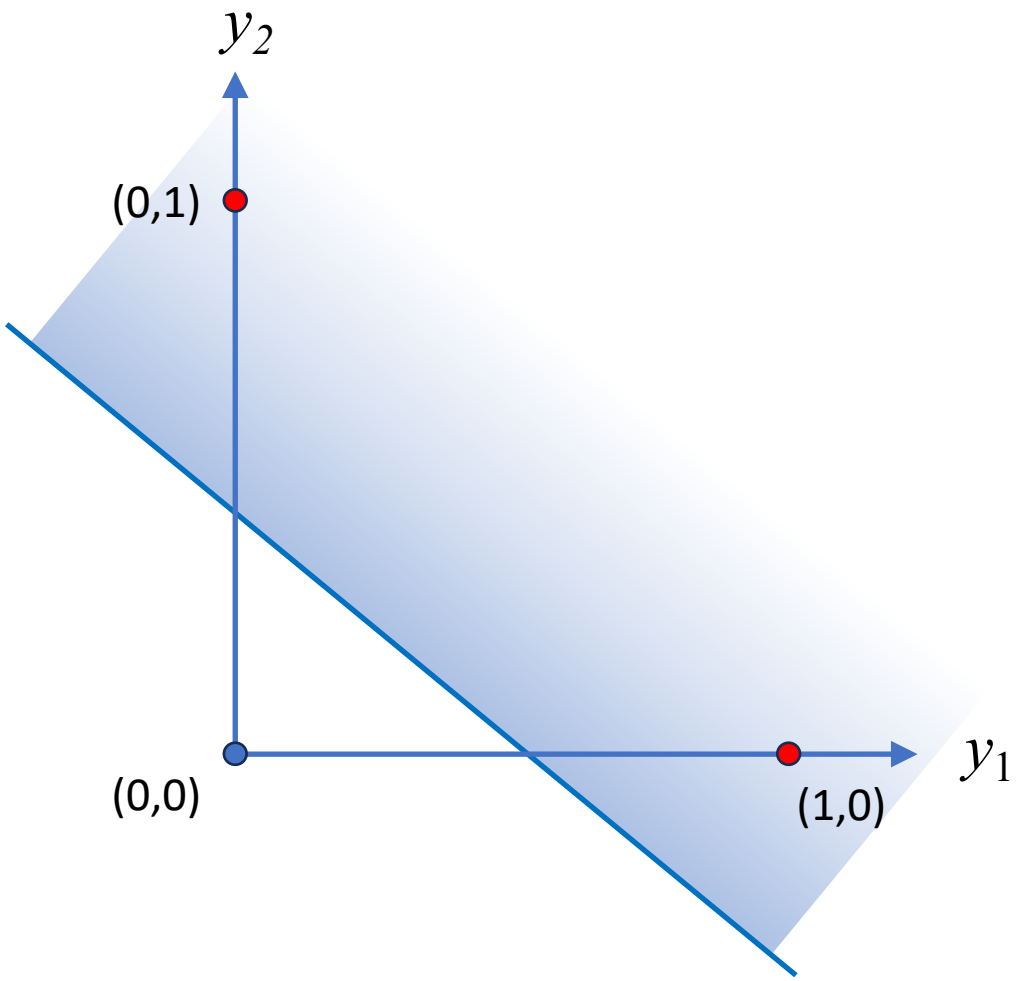


The neural network has a layer that produces a simpler representation in a transformed space where class separation can be performed with a single additional neuron.

Two-neuron neural network

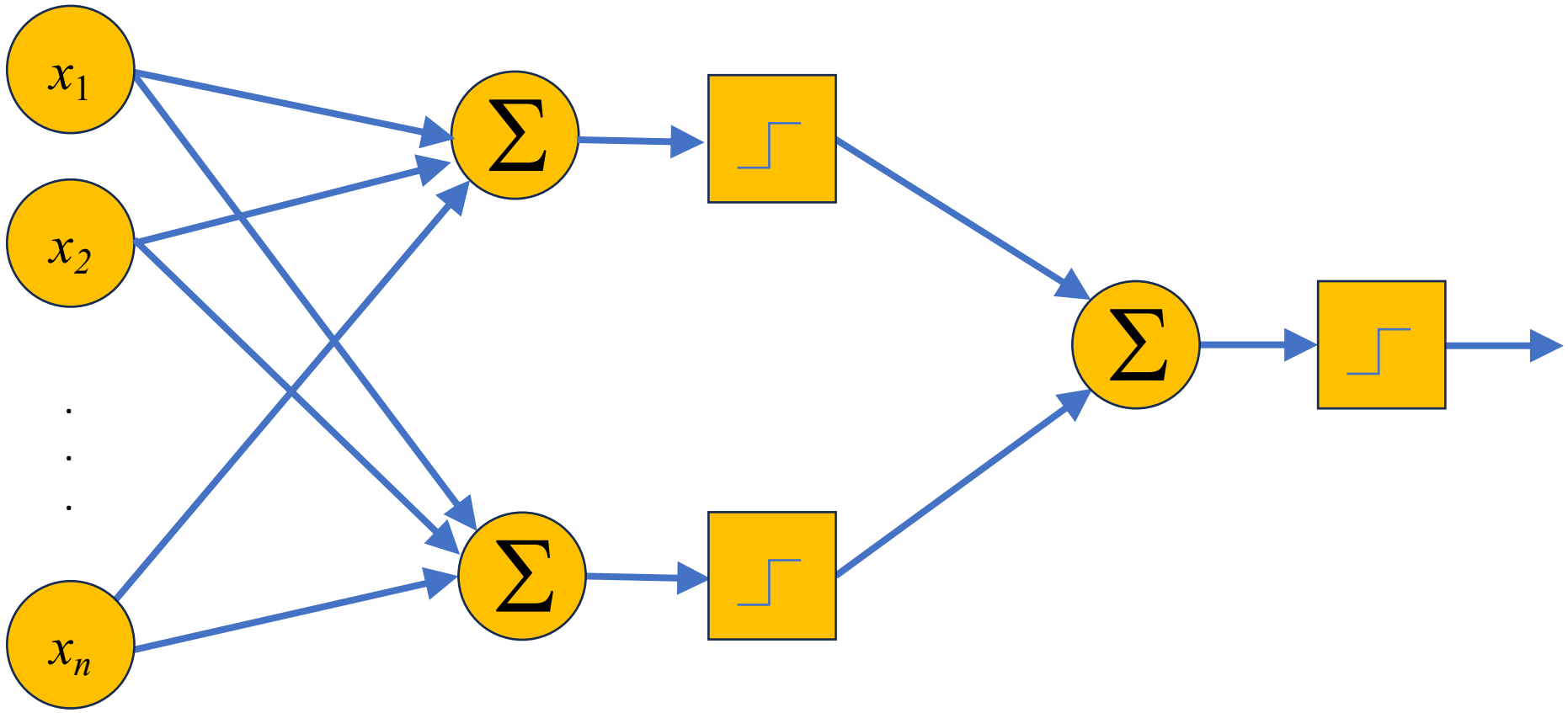
This simple network reproduces the XOR function





n-input neural network

larger and better performing



The dual nature of neural networks

A neural network is defined by:

a) its physical structure (the number of neurons in each layer; the connections between neurons)

b) the set of weights W_i

A neural network does not perform as expected without defining the proper set of weights:

c) the weights are learned from a set of examples

d) the network can be used after the learning step

Prior distributions

The choice of prior distribution is an important aspect of Bayesian inference

- prior distributions are one of the main targets of frequentists: how much do posteriors differ when we choose different priors?
- there are two main “objective” methods for the choice of priors (MaxEnt and Jeffreys')
- here we discuss

1. The quest for "objective" priors
2. Review of the Cramer-Rao bound and related concepts
3. Information-theoretic concepts in statistics
4. Jeffreys' method
5. Reference priors
6. The Maximum Entropy Method

Random variable transformations and prior distributions

$$p_x(x)dx = p_x\left[x(y)\right]\left|\frac{dx}{dy}\right|dy = p_y(y)dy$$

$$\Rightarrow \quad p_y(y) = p_x\left[x(y)\right]\left|\frac{dx}{dy}\right|$$

- In general, if the first pdf is uniform, the other one is not. This means that choosing a uniform distribution as the "least informative" distribution is not enough, *unless we specify which variate should be uniformly distributed*.

- How can we "objectively" choose a prior distribution???

Review of the Cramer-Rao bound - proof of the Bartlett identities

- pdf normalization

$$\int_{\{x\}} p(x,\theta)dx = 1$$

- derivation of normalization formula

$$\frac{\partial}{\partial\theta}\int_{\{x\}} p(x,\theta)dx = \int_{\{x\}} \frac{\partial p(x,\theta)}{\partial\theta}dx = 0$$

- further manipulation of the previous result

$$\begin{aligned} 0 &= \int_{\{x\}} \frac{\partial p(x,\theta)}{\partial\theta}dx = \int_{\{x\}} \frac{1}{p(x,\theta)}\frac{\partial p(x,\theta)}{\partial\theta}p(x,\theta)dx \\ &= \int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial\theta}p(x,\theta)dx \\ &= \mathbb{E}\left[\frac{\partial \ln p(x,\theta)}{\partial\theta}\right] \end{aligned}$$

First Bartlett identity

Review of the Cramer-Rao bound - proof of the Bartlett identities (ctd.)

- derivation of the first identity

$$\int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx = 0$$

- further manipulation of the previous result

$$\begin{aligned}
0 &= \frac{\partial}{\partial \theta} \int_{\{x\}} \frac{\partial \ln p(x,\theta)}{\partial \theta} p(x,\theta) dx = \int_{\{x\}} \left[\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} p(x,\theta) + \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial p(x,\theta)}{\partial \theta} \right] dx \\
&= \int_{\{x\}} \left[\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} + \frac{1}{p(x,\theta)} \frac{\partial \ln p(x,\theta)}{\partial \theta} \frac{\partial p(x,\theta)}{\partial \theta} \right] p(x,\theta) dx \\
&= \int_{\{x\}} \left\{ \frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} + \left[\frac{\partial \ln p(x,\theta)}{\partial \theta} \right]^2 \right\} p(x,\theta) dx \\
&= \mathbb{E}\left[\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2} \right] + \mathbb{E}\left[\left(\frac{\partial \ln p(x,\theta)}{\partial \theta} \right)^2 \right]
\end{aligned}$$

Second Bartlett identity

The Cramér-Rao bound and the Fisher information

- using both identities $$\mathrm{var}\left[\frac{\partial \ln L(x,\theta)}{\partial\theta}\right] = \mathbb{E}\left[\left(\frac{\partial \ln L(x,\theta)}{\partial\theta}\right)^2\right] = -\mathbb{E}\frac{\partial^2 \ln L(x,\theta)}{\partial\theta^2}$$

- expectation value of ML estimator $$\mathbb{E}[\hat{\theta}(D)] = \int_{\{x\}} \hat{\theta}(x) L(x,\theta) dx = \theta + b_n$$

- derivative of the previous expression

$$\frac{\partial}{\partial\theta}\mathbb{E}[\hat{\theta}(D)] = 1 + \frac{\partial b_n}{\partial\theta} = \int_{\{x\}} \hat{\theta}(x)\frac{\partial L(x,\theta)}{\partial\theta} dx = \int_{\{x\}} \hat{\theta}(x)\frac{\partial \ln L(x,\theta)}{\partial\theta} L(x,\theta) dx$$

$$= \mathbb{E}\left[\hat{\theta}(D)\frac{\partial \ln L(D,\theta)}{\partial\theta}\right] = \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial\theta}\right] + \mathbb{E}\left[\hat{\theta}(D)\right]\mathbb{E}\left[\frac{\partial \ln L(D,\theta)}{\partial\theta}\right]$$

$$= \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial\theta}\right]$$

The Cramér-Rao bound and the Fisher information (ctd.)

- we use Schwartz's inequality for covariance $[\mathrm{cov}(x,y)]^2 \leq \sigma_x^2 \sigma_y^2$
- we apply the inequality to the previous result $1 + \frac{\partial b_n}{\partial \theta} = \mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right]$ and find

$$\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2 = \left\{\mathrm{cov}\left[\hat{\theta}(D), \frac{\partial \ln L(D,\theta)}{\partial \theta}\right]\right\}^2 \leq \mathrm{var}[\hat{\theta}(D)]\mathrm{var}\left[\frac{\partial \ln L(D,\theta)}{\partial \theta}\right]$$

- rearranging terms, we obtain the **Cramér-Rao bound**

$$\mathrm{var}[\hat{\theta}(D)] \geq \frac{\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2}{\mathrm{var}\left[\frac{\partial \ln L(D,\theta)}{\partial \theta}\right]} = \frac{\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2}{\mathbb{E}\left[\left(\frac{\partial \ln L(D,\theta)}{\partial \theta}\right)^2\right]} = \frac{\left(1 + \frac{\partial b_n}{\partial \theta}\right)^2}{-\mathbb{E}\frac{\partial^2 \ln L(D,\theta)}{\partial \theta^2}}$$

Definition of Fisher Information. A very concentrated pdf is very informative. Therefore, the smaller the variance, the greater the "information".

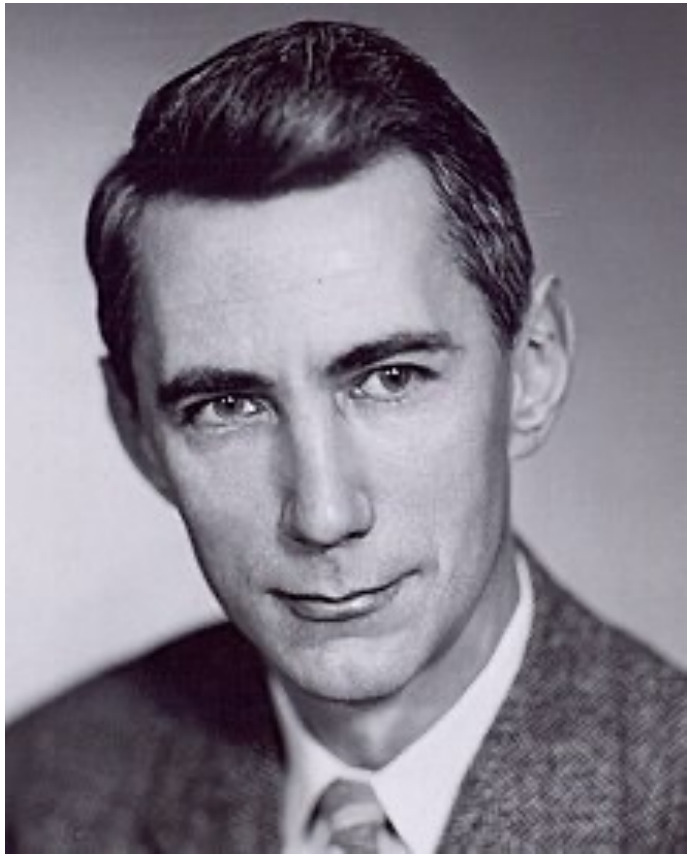
Thus, from the (unbiased, consistent) Cramér-Rao bound

$$\mathrm{var}[\hat{\theta}(D)] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \ln L(D,\theta_0)}{\partial \theta_0}\right)^2\right]} = \frac{1}{-\mathbb{E}\frac{\partial^2 \ln L(D,\theta_0)}{\partial \theta_0^2}}$$

one is led to the Fisher Information

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ln p(x,\theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\frac{\partial^2 \ln p(x,\theta)}{\partial \theta^2}$$

Information theoretic concepts in statistics



Claude Shannon
1916–2001

After graduating from Michigan and MIT, Shannon joined the AT&T Bell Telephone laboratories in 1941.

His paper ‘A Mathematical Theory of Communication’ published in the Bell System Technical Journal in 1948 laid the foundations for modern information theory. This paper introduced the word ‘bit’, and his concept that information could be sent as a stream of 1s and 0s paved the way for the communications revolution.

It is said that von Neumann recommended to Shannon that he use the term entropy, not only because of its similarity to the quantity used in physics, but also because “nobody knows what entropy really is, so in any discussion you will always have an advantage”.

Information theoretic concepts in statistics

- The amount of information can be viewed as a "degree of surprise" on learning the value of a variable *x*. If we are told that a highly improbable event has just occurred, we will have received more information than if we were told that some very likely event has just occurred, and if we knew that the event was certain to happen we would receive no information.

- Thus, information carried by an event (symbol) must be a function of its probability

- If two events (symbols) are independent the total information is the sum of the information carried by each of them, therefore

$$p(x)p(y) \quad \Rightarrow \quad h(x) + h(y)$$

- logarithms have this property, therefore we take

$$h(x) \propto \ln p(x)$$

- more specifically, we choose

$$h(x) = -\log_2 p(x)$$

Information theoretic concepts in statistics – 2

- definition of the **Shannon entropy: average information** carried by the events (symbols)

$$H = -\sum_{k=1}^{N} p_k \log_2 p_k = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k}$$

Example:

- just two symbols, 0 and 1, same probability

$$H = -2\left(\frac{1}{2}\log_2 \frac{1}{2}\right) = 1 \text{ bit}$$

there are 2 equal terms

average information conveyed by each symbol

the result is given in pseudounit "bits" (for natural logarithms this is "nats")

Information theoretic concepts in statistics – 2

- just two symbols, 0 and 1, probabilities ¼ and ¾ , respectively

$$H = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} \approx 0.81 \text{ bit}$$

- 8 symbols, equal probabilities

$$H = -\sum_1^8 \frac{1}{8}\log_2\frac{1}{8} = \log_2 8 = 3 \text{ bit}$$

- 8 symbols, with probabilities ½, ¼, 1/8, 1/16, 1/64, 1/64, 1/64, 1/64

$$\begin{aligned} H &= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{16}\log_2\frac{1}{16} + 4\times\frac{1}{64}\log_2\frac{1}{64}\right) \\ &= \frac{1}{2} + \frac{1}{2} + \frac{3}{8} + \frac{1}{4} + 4\times\frac{3}{32} = 2 \text{ bit} \end{aligned}$$

Information theoretic concepts in statistics – entropy in statistical mechanics

- consider a system where states *n* are occupied by N_n identical particles (*n, n=1, … , M*).

- the number of ways to fill these states is given by

$$\Omega = \frac{N!}{N_1!N_2!\dots N_M!}$$

- then Boltzmann's entropy is

$$H_B = k_B \ln \Omega = k_B \ln \frac{N!}{N_1!N_2!\dots N_M!} \approx k_B \left[(N \ln N - N) - \sum_i (N_i \ln N_i - N_i) \right]$$

$$= k_B \left[N \ln N - \sum_i N p_i (\ln p_i + \ln N) \right] = k_B N \sum_i p_i \ln \frac{1}{p_i}$$