UAB

Modelling for Science and Engineering

Is there a Science for AI? Part I

S. Xambó

IMTech & BSC

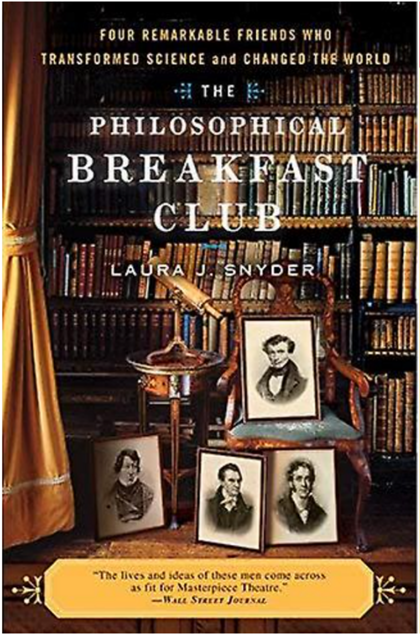
23/1/2023

N

Preface

On Science · Human-machine dialectics

https://web.mat.upc.edu/sebastia.xambo/99/s-uab.pdf


FOUR REMARKABLE FRIENDS WHO
TRANSFORMED SCIENCE and CHANGED THE WORLD
THE
PHILOSOPHICAL
BREAKFAST
CLUB
LAURA J. SNYDER
"The lives and ideas of these men come across
as fit for Masterpiece Theatre."
—Wall Street Journal


ACANTILADO
Laura J. Snyder
El Club de los
desayunos filosóficos
Cuatro notables amigos que transformaron
la ciencia y cambiaron el mundo
TRADUCCIÓN DE JOSÉ MANUEL ÁLVAREZ-FLÓREZ

William Whewell (1794-1866). Publised the epoch-making treatise *The Philosophy of the Inductive Sciences, Founded upon Their History*, in 2 volumes (1840).

Influenced many researchers, as for example Michael Faraday, Charles Darwin, Charles Lyell, John Herschell, Richard Jones, etc.

Founded the fields of crystallography, mathematical economics, and the science of tides. Coined the words scientist, anode, cathode, and ion.

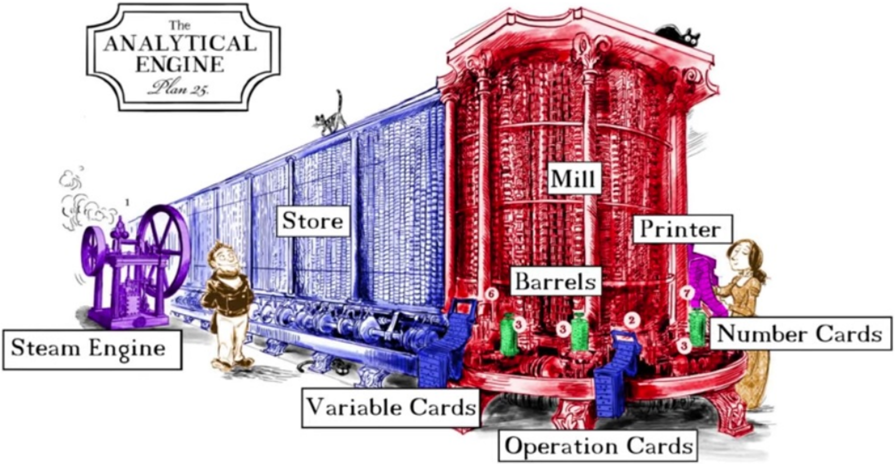
They were at the vanguard of the modernization of science: Herschel mapped the skies of the Southern Hemisphere and contributed to the invention of photography; Jones shaped the science of economics; Babbage invented the analytical engine, which embodied the main ideas of the modern computer.

Charles Babbage (1792-1871): A mathematician, philosopher, inventor and mechanical engineer, Babbage originated the concept of a digital programmable computer (WP).

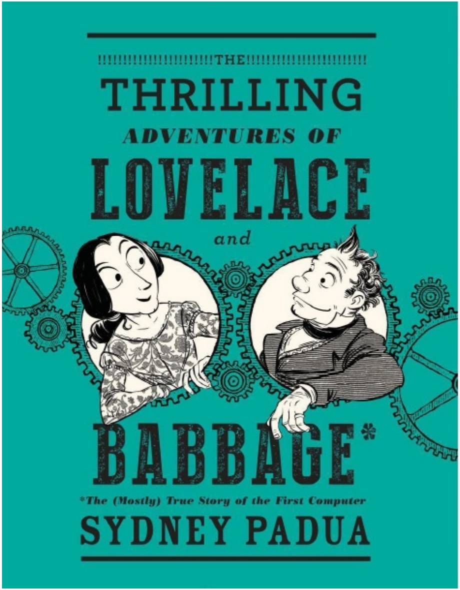
No instance of his **analytical engine** has ever been constructed, but there is work in progress to build one according to Babbage's Plan 28.

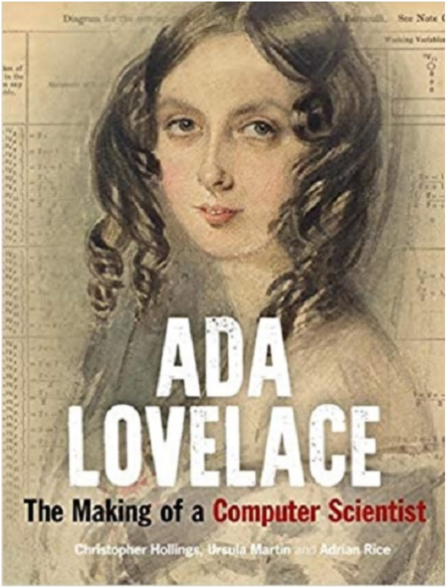
He calculated a table of logarithms with his earlier **difference engine**, an important task at the time that had been suffering from errors incurred by the human calculators.

[1] (collier-maclachlan-2000)



From *Charles Babbage's Analytical Engine* (video by Michael Holzheu, illustrations by Sidney Padua). See also *Babbage's Analytical Engine* (talk by David F. Brailsford at the December 2016 Ada Lovelace Symposium), and Plan28.

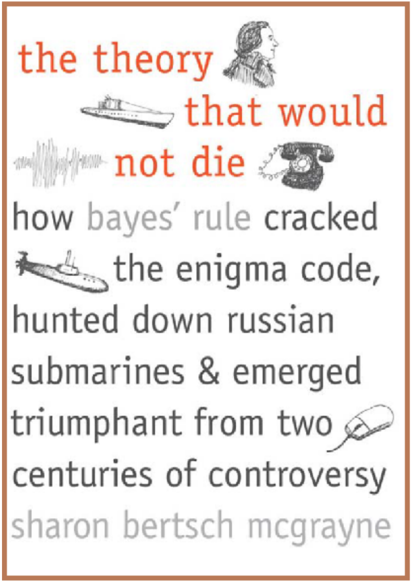




[2] (padua-2015) [3] (hollings-martin-rice-2018)

Chances and Logic

The Bayes-Laplace rule
Two faces of entropy
Homage to Turing



[4] (mcgrayne-2011) N [5] (silver-2012)

Learning from experience is neatly encoded, in probabilistic terms, by the *Bayes-Laplace rule*:

$$P(x|y) = P(x)P(y|x)/P(y) = P(x)K(x,y),$$

where $K(x,y) = P(y|x)/P(y) = P(x \cap y)/P(x)P(y)$ is symmetric.

This tells us how to update our *prior* belief in x, $P(x)$, to the belief $P(x|y)$ *posterior* to having observed y: $P(y|x)$ is the *likelihood* of y, and $P(y)$ is the *evidence* of y. When $K > 1$ ($K < 1$), our believe is *increased* (*decreased*), and in any case it may be construed as a *learning* on x produced by the observation of y.

The Bayes–Laplace rule is the basis of many sorts of AI models (we provide a few references on page 19 and on page 23).

In practice, the trouble arises from the evaluation of the *evidence* $P(y)$ of y, as $P(y) = \sum_x P(y|x)P(x)$, and this sum (or integral in continuous models) is intractable but for the simplest situations.

Let $\{i\}$ be a set of indices for the internal states of a system composed of a large number of particles. Let E_i denote the energy of the system in the state i. If we can assume that the system is in thermal equilibrium with its surrounding environment, then the probability that the system is in state i is given by the *Gibbs distribution*:

$$p_i = \tfrac{1}{Z} e^{-E_i/kT},$$

where Z is a constant independent of all states, T is the absolute temperature (in kelvins) and $k = k_B = 1.38 \times 10^{-23}$joule/kelvin is *Boltzmann's constant*. Note that $\sum p_i = 1$ implies what is called the *partition function*:

$$Z = \sum_i e^{-E_i/kT}$$

From the expression giving p_i and the properties of the exponential function, we see that p_i increases when E_i decreases (so low energy states are more probable than high energy states) and that the probability of the low energy states increases when T decreases.

Note also that

$$\log p_i + \log Z = -E_i/kT, \quad \text{or} \quad kT\log p_i + kT\log Z = -E_i.$$

If we multiply the last relation by p_i and sum over i, we get:

$$-TS + kT\log Z = -\langle E\rangle,$$

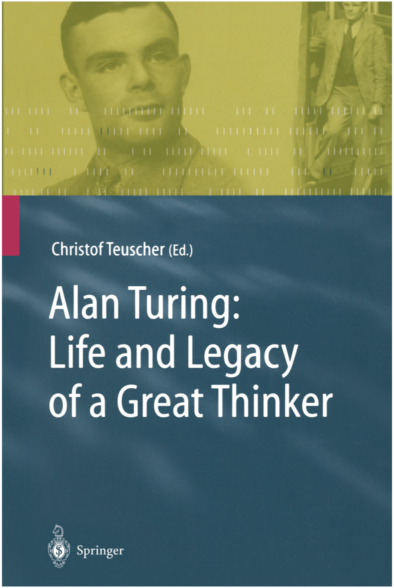
where $S = -k\sum_i p_i \log p_i$ (*Gibbs entropy*) and $\langle E\rangle = \sum_i E_i$ (*average energy*). Thus we conclude that

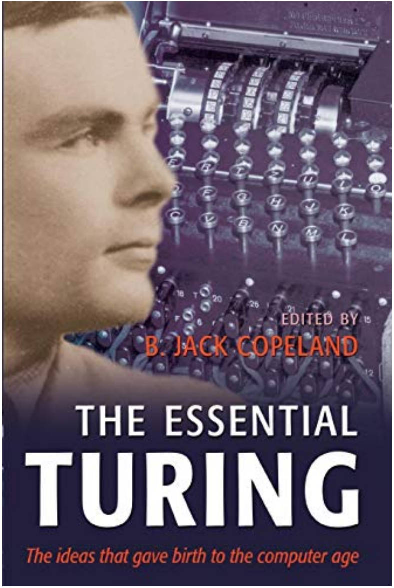
$$\langle E\rangle = TS + F, \quad \text{or} \quad F = \langle E\rangle - TS, \qquad (*)$$

where $F = -kT\log Z$ is the (Helmholtz) *free energy*. But $(*)$ is the classical thermodynamical relation defining *entropy* (S), so we can conclude that *the classical entropy is equal to the Gibbs entropy, which itself is proportional to Shannon's entropy*. This explains the deep reason for why Shannon chose the word entropy to name the quantity $-\sum_i p_i \log p_i$.

(adapted from [6, Section 11.2])

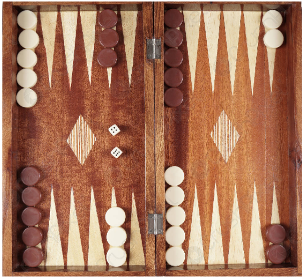
N

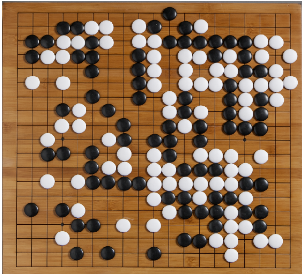


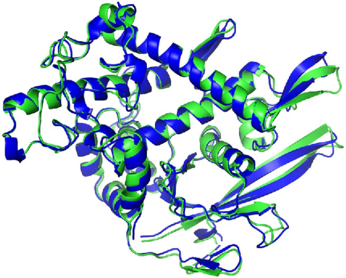


[7] (teuscher-2004) N [8] (copeland-2004)

AIs are blowing in the wind







Chess, Backgammon, Go, Console games, Protein folding, ...

Beethoven's Unfinished 10th Symphony Brought to Life by Artificial Intelligence (Scientific American, October 15, 2021).
See Classic *f*M. Sample.

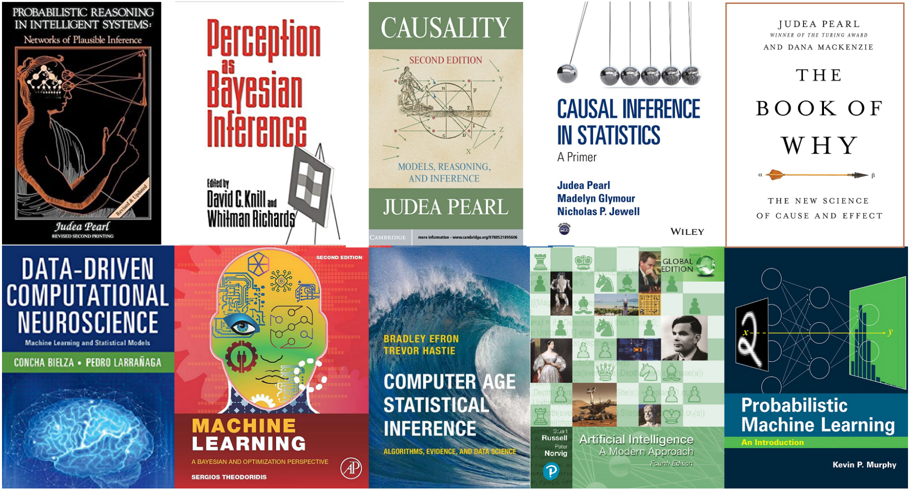
Walter Werzowa: "I dare to say that nobody knows Beethoven as well as the AI did—as well as the algorithm". "I think music, when you hear it, when you feel it, when you close your eyes, it does something to your body. Close your eyes, sit back and be open for it, and I would love to hear what you felt after".

We will see more examples later.


Neural Networks and Learning Machines
Third Edition
Simon Haykin
Shai Shalev-Shwartz and Shai Ben-David
UNDERSTANDING MACHINE LEARNING
FROM THEORY TO ALGORITHMS
DEEP LEARNING
Ian Goodfellow, Yoshua Bengio, and Aaron Courville
Foundations of Machine Learning second edition
Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar
Gopinath Rebala · Ajay Ravi
Sanjay Churiwala
An Introduction to Machine Learning
Springer
INTRODUCTION TO
Machine Learning
fourth edition
ETHEM ALPAYDIN
Learning Theory from First Principles
DRAFT
October 13, 2022
Francis Bach
francis.bach@inria.fr
THE PRINCIPLES OF DEEP LEARNING THEORY
An Effective Theory Approach to Understanding Neural Networks
Daniel A. Roberts and Sho Yaida
based on research in collaboration with Boris Hanin

General

N



Bayesian approaches

N

Applications

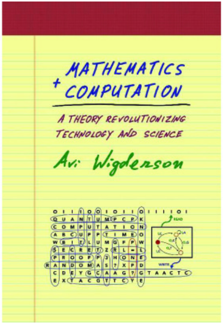
Many facets of the symbiosis *Mathematics & Computation* are presented in [9] (wigderson-2019).

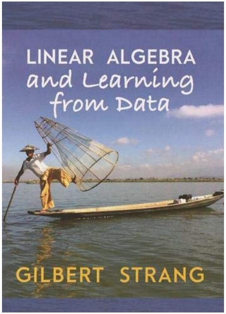
In particular, Chapter 17 is devoted to *computational learning theory*.

Another useful reference is [10] (strang-2019).

See also the extensive *survey* [11] (nguyen-dlugolinsky-bobak-et-5-2019) and the popular approach [12] (Marr's blog).

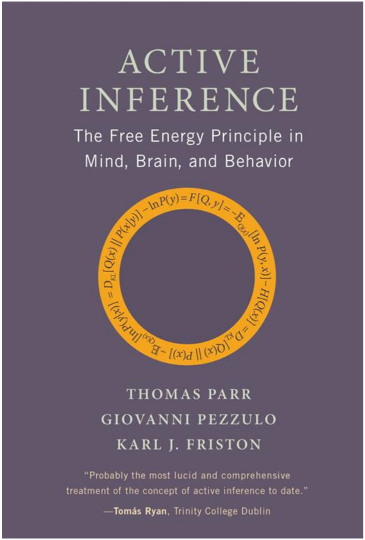
And, of course, articles in the Wikipedia, such as `Machine_learning`, keep growing in number and substance.





How do they work?

Bayesian brain
Clustering methods
Supervised learning
Reinforcement learning



[13] (parr-pezzulo-friston-2022)

- $H(Q(x))$: $-\sum_x Q(x)\ln Q(x)$ (*Entropy* of $Q(x)$)
- $KL(Q(x)||P(x))$: $\sum_x Q(x)\ln\frac{P(x)}{Q(x)}$ (*Kullback-Leibler divergence*)
- $F(Q,y)$: *Free energy*

The golden ring

$$
\begin{aligned}
F(Q,y) &= \underbrace{-\mathbb{E}_{Q(x)}[\ln P(y,x)]}_{Energy} - \underbrace{H[Q(x)]}_{Entropy} \\
&= \underbrace{KL[Q(x)||P(x)]}_{Complexity} - \underbrace{\mathbb{E}_{Q(x)}[\ln P(y|x)]}_{Accuracy} \\
&= \underbrace{KL[Q(x)||P(x|y)]}_{Divergence} - \underbrace{\ln P(y)}_{Evidence}
\end{aligned}
$$

A shorter account is provided by the tutorial paper [14] (smith-friston-whyte-2022) on "active inference and its application to empirical data". N

Aimed at finding *hidden structure in data*, $\mathcal{D} = \{x^1, \ldots, x^m\}$.

***k*-Means.** This algorithm groups unlabeled data $\mathcal{D}$ in k classes:

(1) Select k vectors $z^1, \cdots, z^k \in \mathcal{D}$ at random.

(2) Assign each $x^j \in \mathcal{D}$ to the first z^i nearest to x^j (initial groups).

(3) Update each z^i to the centroid (or *mean*) of the z^i group.

(4) Iterate (2) and (3) until the z^i are stable (up to a *tolerance*).

The associated *cluster predictor* assigns x to the first nearest z^i.

***k*-NN** (nearest neighbors). Let $\mathcal{D} = \{(x^1, y^1), \cdots, (x^m, y^m)\}$ be a labeled set and k a positive integer. The *label predictor* of the k-NN algorithm assigns a vector x to the mode of $y^{j_1}, \cdots, y^{j_k}$, where $x^{j_1}, \cdots, x^{j_k}$ are the nearest neighbors of x from among $x^1, \cdots, x^m$.

Hypothesis space[3]

$$\mathcal{H} = \{h\}$$

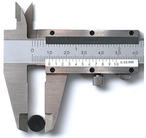
Data, expert tables[2]

$$\mathcal{D}$$
$$\{(x^1, y^1), \ldots, (x^m, y^m)\}$$
$$x^j \sim P, \quad y^j = f^*(x^j)$$

World – Observer[1]

Learner/LA[5]

Error measure[4]



$\Longrightarrow$

Prediction rule[6]

$$y = h(x)$$

[1] *Urbi et Orbi* engraving (Flammarion). Observer produces data (**Tycho Brahe**).
[2] Tables of planet and comet positions over time (**Ephemeris**)
[3] Hypothesis space (*Inductive bias*). Greeks/Copernicus: circles around the Earth/Sun.
[4] Error measure (*loss*, *risk*, *regret*): How close are predictions to observations?
[5] Learning algorithm (**Kepler**). New bias: Ellipses with a focus at the Sun.
[6] The algorithm supplies a predicton h. Hopefully, $h \approx f^*$.

N

Set of data $x^j \in \mathbf{R}^n$, $j \in [m]$ (*dataset*).

Want to predict values y^j provided by a *supervisor* or *expert* in such a way that for objects x not in the dataset the value y corresponding to x is predicted with high probability (*generalization* capacity).

Hypothesis space: a space of parameterized functions,

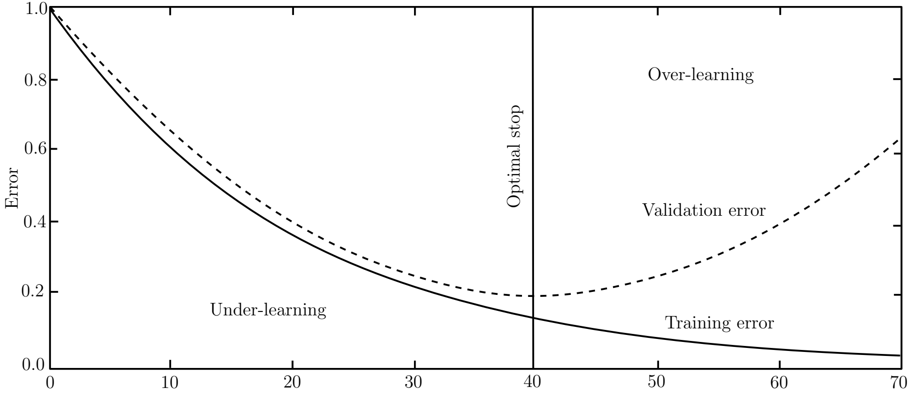
$$\mathcal{H} = \{h_w\}_{w \in W}.$$

Problem: to find $w \in W$ such that $h_w(x^j) \approx y^j$.

Method: If the criterion for fitness depends on a function $\ell(h(x), y)$ (*local cost*), the problem amounts to finding w that minimizes $\sum_j \ell(f_w(x^j), y^j)$ (*empirical cost* or *risk*):

$$\mathrm{argmin}_w \sum_j \ell(h_w(x^j), y^j).$$

This is the *empirical risk minimization* rule, ERM.



Scheme of a basic training process. In one *epoch* (one step in the training loop), f_w is applied to the dataset. The proportion of errors is the *training error* for that epoch. The learning algorithm determines an update $w = w - \Delta w$ (for example with some variant of gradient descent) and a new epoch is run. The *validation error* is the error incurred by f_w on a *validation dataset*. The training error is decreasing, while the validation error decreases until the *optimal stop* and thereafter it increases.

Supervised learning has two main modalities:

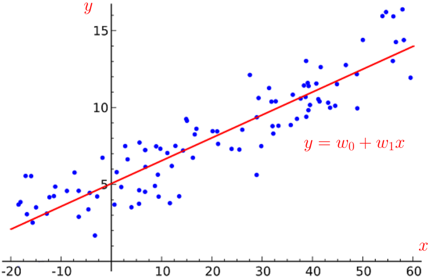
Classification: When the set $\mathcal{Y}$ is finite. In this case its elements are usually called *labels* or *classes*.

Regression: When the set $\mathcal{Y}$ is the set $\mathbf{R}$ of real numbers.

Example. In *linear regression*, $\mathcal{H}$ is the space of functions of the form

$$h(x) = w_0 + w_1 x_1 + \cdots + w_n x_n$$

and the local cost is usually $(h(x) - y)^2$.



$\mathcal{D} = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ $(x^j, y^j \in \mathbf{R})$.

Problem. Find a polynomial map $p : \mathbf{R} \to \mathbf{R}$ of degree r,

$$p(x) = w_0 + w_1 x + \cdots w_r \cdot x^r, \quad w_0, w_1, \cdots, w_r \in \mathbf{R},$$

such that $\hat{y}^j = p(x^j)$ are as close as possible to the y^j (*polynomial approximation of degree r*).

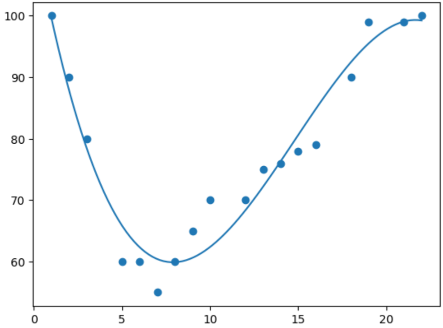


Figure 8.1: Cubic approximation of a dataset in $\mathbf{R} \times \mathbf{R}$.

The *logistic regression* is linear regression of $\log p/(1-p)$.

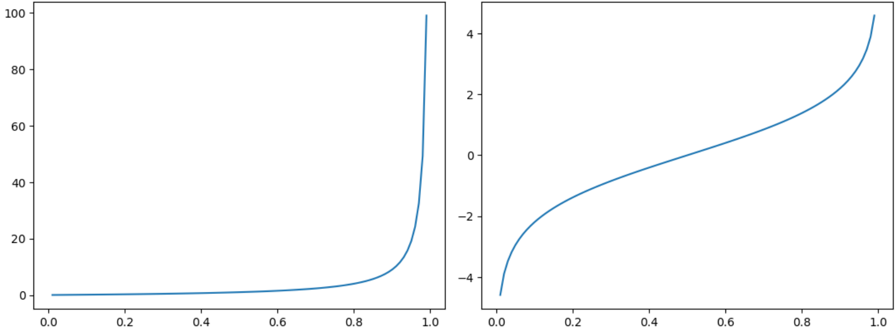
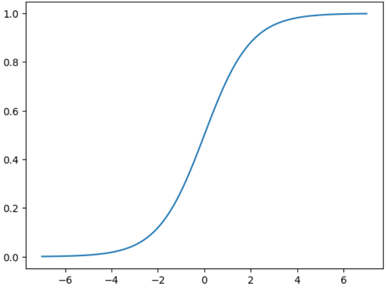


Figure 8.2: For probability values $p \in [0,1]$ it makes no sense to apply linear regression procedures. Left: graph of the *odds* function, $p/(1-p)$, for $p \in [0,1]$. Right: graph of $\log(p/1-p)$, with symmetry about the point $(1/2, 0)$, so linear regression of its values is in principle possible.

If $\log(p/1-p) = w \cdot x$ ($x, w \in \mathbf{R}^n$), then $p = p(x) = 1/(1+e^{-w\cdot x})$ estimates the probability of observations x.

The function $1/(1+e^{-t})$ is the *logistic function*. Its range is $(0,1)$. A variation is the function $(1-e^{-t})/(1+e^{-t})$, with range $(-1,1)$.



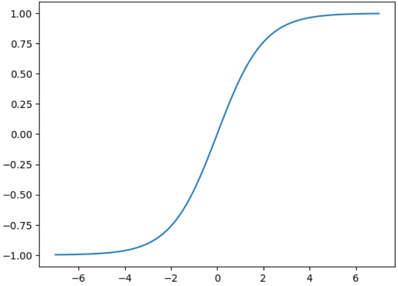
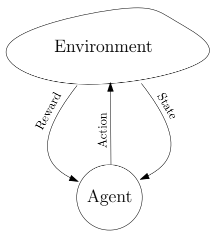
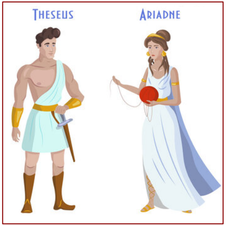
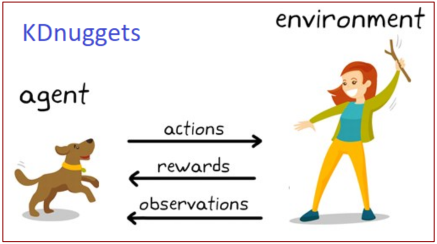


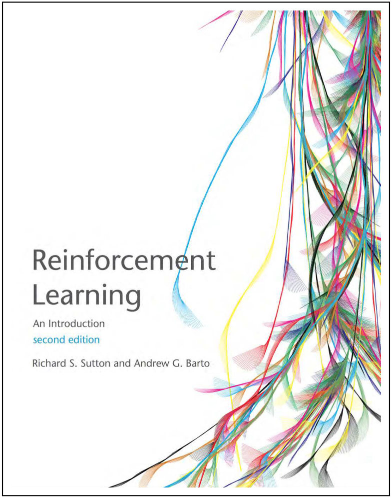
Figure 8.3: Logistic (or *sigmoid*) functions.

Algorithm learns to *react* to an *environment* that provides *rewards*









[15] (sutton-barto-2018)

N

End of Part I

Notes

Abstract

After some preliminary considerations about the historical unfolding of Science, and then about the development of the multiform field of Artificial Intelligence, we will delve into some of the strands (like algorithmic learning, Bayesian networks, various flavours of neuroscience, or statistical physics) that seem more likely to play a key role in building a Science able to account for biological and artificial intelligences. Such a Science ought to grant understanding of experimental work, foster novel algorithmic insights, and supply guidance for innovative engineering designs. Along the way we will have a closer look on a fair sample of contemporary references and leading researchers with the potential to inspire and orient people wishing to contribute to this Promethean human endeavour. P

The Philosophical Breakfast Club recounts the life and work of four men who met as students at Cambridge University: Charles Babbage, John Herschel, William Whewell, and Richard Jones. Recognizing that they shared a love of science (as well as good food and drink) they began to meet on Sunday mornings to talk about the state of science in Britain and the world at large. Inspired by the great 17th century scientific reformer and political figure Francis Bacon—another former student of Cambridge—the Philosophical Breakfast Club plotted to bring about a new scientific revolution. And to a remarkable extent, they succeeded, even in ways they never intended.

P

A general-purpose, programmable, steam-powered computing machine that incorporated arithmetic logic, conditional branching, loops and integrated memory, and would have used punch-cards and a printer to input and output data. P

This image was the entry icon of the AI section included in the exhibit Imaginary-2021 held at the FME in the Fall of 2021.

It symbolizes the dialectics between two approaches to model intelligent behaviors. One, represented by the artificial hand, relies on *computational sciences* in a broad sense. The other, represented by the human hand, relies on *brain sciences*, also in a broad sense.

In my talk I will consider both approaches, not as separate domains, but rather as *intertwined disciplines*. This allows a sort of *in crescendo* unfolding that will culminate with a glimpse over possible syntheses of the two. All along, I will try to emphasize the two-way bridges with mathematics. I will also provide a good number of reading and study references. P

The first, Sharon Mcgrayne's book that you see on the screen. A well-deserved praise for the Bayes-Laplace rule that I will discuss below. Let me add an extraordinary book by Nate Silver: *The signal and the noise–Why so many predictions fail, but some don't* (Penguin Press 2012). P

A memory also for Alan Turing, the initiator of so many important things, in the form of two books. They are very good, but the first only deals superficially with his ideas about morphogenesis and artificial life, while in the second the last three chapters are devoted to these topics. Surely you have noticed that the cover of McGrayne's book contains the clause "how bayes' rule cracked the enigma code", one of Turing's most spectacular practical triumphs. P

Brandon Walker, 2020. Games that AI have played have thus captured the interest of news, here are the most important victories it has won.
1996–1997 Kasparov v. IBM Deep Blue.
2011 IBM Watson Wins Jeopardy!
2013 DeepMind Beats Atari.
2016 AlphaGo v. Lee Sedol.
2017 AlphaZero Masters Chess, Go, and Shogi.
2019 AlphaStar to play Starcraft.

In February 1996 Kasparov won Deep Blue 4-2 (lost match 1, won matches 2,5,6, tie 3,4). In May 1997 (3-11), IBM's Deeper Blue wins Kasparov $3\frac{1}{2} - 2\frac{1}{2}$ (won 1, lost 2, 6, draw 3, 4, 5. Many materials for reflection: the 2003 movie Game Over, and Kasparov's 2017 book (with Greengard) *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*, [16].

AlphaGo (2016). After defeating the European Go champion Fan Hui, AlphaGo was tasked with playing Lee Sedol, one of the highest ranking players ever. Beating Lee prompted him to retire, saying "Even if I become the number one, there is an entity that cannot be defeated".

AlphaZero (2017) was a generalized version of AlphaGo, built with the intention of winning Chess, Go and Shogi (a Japanese version of chess). Not only did AlphaZero beat AlphaGo, it was *done by only playing simulated games against itself, having no examples of expert's games to look at*. *At the start of these simulations it knew absolutely nothing*. *It mastered Chess after 9 hours of training, Shogi after 2, and Go after 34*.

AlphaStar. AlphaZero was then transitioned into AlphaStar with the intention of beating the real-time strategy game Starcraft. In 2019 AlphaStar achieved a ranking in the top 0.2 percent of human players. This was the first time that an AI had ever topped an e-sport.

"Starcraft," released in 1998 by Blizzard Entertainment, is a real-time strategy game where players build a military base, mine resources, and attack other bases.

Other games introduce demands of handling incomplete information (e.g. StarCraft), understanding narrative (e.g. Skyrim), or very long-term planning (e.g. Civilization).

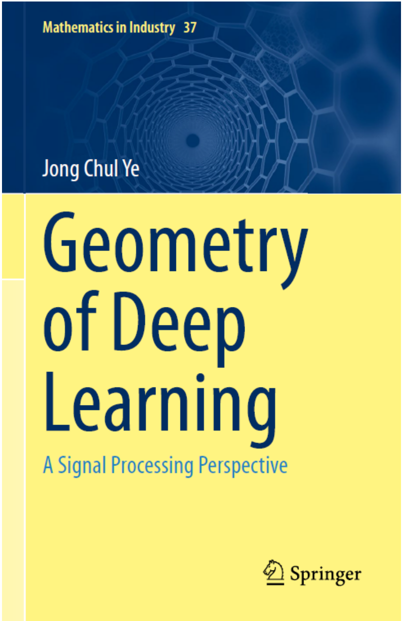
More info: [17] (general), [18] and [19] (backgammon), [20] (Go). See also [16] (an account by Kasparov about his 'chess defeat' by IBM Deep Blue). P

General

[6] (haykin-2009)
[21] (shalevshwartz-bendavid-2014)
[22] (goodfellow-bengio-courville-2016)
[23] (mohri-rostamizadeh-talwalkar-2018)

[24] (rebala-ravi-churiwala-2019)
[25] (alpaydin-2020)
[26] (bach-2022)
[27] (roberts-yaida-hanin-2022)

P



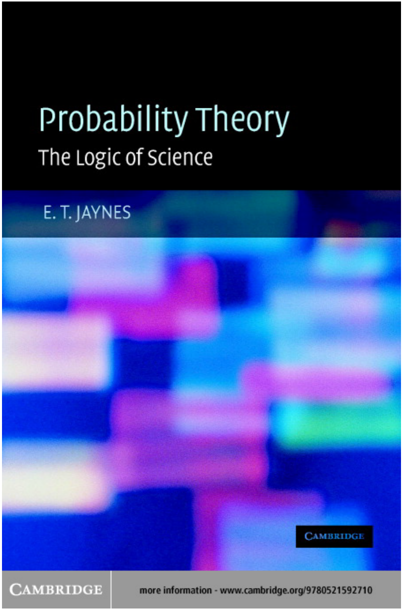
[28] (ye-2022)

Bayesian approaches

[29] (pearl-1988)
[30] (knill-richards-1996)
[31] (pearl-2009)
[32] (pearl-glymour-jewell-2016)
[33] (pearl-mackenzie-2018)

[34] (bielza-larranaga-2020)
[35] (theodoridis-2020)
[36] (efron-hastie-2021)
[37] (russell-norvig-2022)
[38] (murphy-2022)

P



[39] (jaynes-2003)

Applications

[40] (guttag-2016)
[41] (nandy-biswas-2018)
[42] (said-torra-2019)
[43] (lavor-X-zaplana-2018)

[44] (balas-roy-sharma-samui-2019)
[45] (chollet-2020)
[46] (colome-torras-2020)
[47] (xambo-2021-iciam)

29 case studies. Example: Supermarket giant Kroger and 84.51°: *138 different machine learning models in production*; over ten petabytes of customer data analyzed; AutoML (makes it possible for those without traditional data science training to create ML models); forecasts for each item in each of more than 2,500 stores for each of the subsequent fourteen days. **Insights**. Example: The platforms that make AI work (209-215). P

[48] (davenport-miller-2022)

Comparing energy and information. The universality and importance of the concept of *information* can be compared with that of *energy*. It is interesting to compare these two (cf. [49]). One is tempted to say that the great inventions of civilization serve either to transform, store and transmit energy (fire, mechanisms like wheels, use of water and wind energy, for instance, for sailing or in mills, steam engines, use of electric, later nuclear energies, rockets, etc.) or they serve to transform, store and transmit information (speech, writing, drum- and fire-signals, printing, telegraph, photograph, telephone, radio, phonograph, film, television, computers, etc.). The analogy goes further. It took a long time (until the middle of the nineteenth century) for the abstract concept of energy to be developed, i.e. for it to be recognized that mechanical energy, heat, chemical energy, electricity, atomic energy, and so on, are different forms of the same substance and that they can be compared, measured with a common measure. What, in fact, remains from the concept of energy, if we disregard its forms of apparition, is its quantity, [its] measure, which was introduced some 125 years ago. In connection with the concept of information, this essentially happened a century later, with the works of Shannon (1948a,b). [There is even a "principle of conservation of information"-like that of energy; see Katona and Tusnády (1967) and Csiszar et al. (1969).] Again, if we disregard the different contents (meanings) of information, what remains is its quantity, [its] measure. From [50, pp. 1-2]. P

Karl Friston (born 12 July 1959): According to his profile in GoogleScholar, he has published more than 500 articles and has more than 300000 citations (consulted 23 October 2022). Fields of work: Neuroscience, Mathematical and theoretical biology, Variational Bayesian methods.

Some recent papers: [51] (sajid-ball-parr-friston-2021), [52] (dacosta-lanillos-sajid-friston-khan-2022), [53] (barp-et2-friston-et3-2022)

K. Friston G. Pezzulo T. Parr

P

Brahe experienced the solar eclipse of 21 August 1560 [he was 15], and was greatly impressed by the fact that it *had been predicted, although the prediction based on current observational data was a day off*. He realized that *more accurate observations* would be **the key to making more exact predictions**.

Kepler relied on Brahe's data to study the orbit of Mars. The circle was discarded, as it produced errors that were incompatible with Brahe's observational precision. Then he tried eccentric circles, and again the errors were two large. Finally he tried ellipses with a focus at the Sun, and with this he succeeded (Kepler's first law). With this model, he could discover how fast a planet moved (in the form of areal velocity, the second law) and to compare the periods of different planets (third law: a_P^3/T_P^2 has the same value for all planets P, hence $a_P \propto \sqrt[3]{T_P^2}$).

P

"Conversely, in *active inference* an agent's interaction with the environment is determined by action sequences that *minimize expected free energy* (and not the expected value of a reward signal). Additionally, unlike in reinforcement learning, the reward signal is not differentiated from other types of sensory outcomes. That is any type of outcome may be more or less preferred. This means that *the implicit reward associated with any outcome is a feature of the creature seeing the observation* – not the environment they inhabit. This may be different for different agents, or even for the same agent at different points in time. This highlights that the two frameworks have fundamentally different objectives: *reward-maximization in reinforcement learning* and *free energy minimization in active inference*" [51] (sajid-ball-parr-friston-2021).

Definition of reinforcement learning in [15] (sutton-barto-2018):

"Reinforcement learning is *learning what to do* – how to map situations to actions – *so as to maximize a numerical reward* signal". P

References I

[1] B. Collier and J. MacLachlan, *Charles Babbage: And the engines of perfection*.

Oxford University Press, 2000.

[2] S. Padua, *The Thrilling Adventures of Lovelace and Babbage: the (mostly) true story of the first computer*.

Pantheon, 2015.

[3] C. Hollings, U. Martin, and A. Rice, *Ada Lovelace: the making of a computer scientist*.

Bodleian Library, Oxford, 2018.

XIII+114 p.

References II

[4] S. B. McGrayne, *The theory that would not die*.

Yale University Press, 2011.

[5] N. Silver, *The signal and the noise: why so many predictions fail–but some don't*.

Penguin, 2012.

[6] S. Haykin, "Neural networks and learning machines," 2009.

xxx + 906 pp.

[7] C. Teuscher, *Alan Turing: Life and legacy of a great thinker*, vol. 54.

Springer, 2004.

Foreword by Douglas Hofstadter.

References III

[8] B. J. Copeland, *The essential Turing*.
Oxford University Press, 2004.
AMS Notices Review,by A. Hodges.

[9] A. Wigderson, *Mathematics and computation*.
Princeton University Press, 2019.
https://www.math.ias.edu/files/mathandcomp.pdf, 372 p.

[10] G. Strang, *Linear algebra and learning from data*.
Wellesley-Cambridge Press, 2019.

References IV

[11] G. Nguyen, S. Dlugolinsky, M. Bobák, V. Tran, Á. L. García, I. Heredia, P. Malík, and L. Hluchỳ, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artificial Intelligence Review*, vol. 52, no. 1, pp. 77–124, 2019.

[12] B. Marr, "The Key Definitions Of Artificial Intelligence (AI) That Explain Its Importance," Consulted 2021.

Key definitions of AI.

[13] T. Parr, G. Pezzulo, and K. J. Friston, *Active inference: the free energy principle in mind, brain, and behavior*.

MIT Press, 2022.

[14] R. Smith, K. J. Friston, and C. J. Whyte, "A step-by-step tutorial on active inference and its application to empirical data," *Journal of mathematical psychology*, vol. 107, p. 102632, 2022.

References V

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction (2nd edition)*.

MIT press, 2018.

[16] G. Kasparov and M. Greengard, *Deep thinking: where machine intelligence ends and human creativity begins*.

Hachette UK, 2017.

[17] J. Togelius, "Better playing through algorithms," 2022.

Togelius blogspot.

[18] G. Tesauro, "Programming backgammon using self-teaching neural nets," *Artificial Intelligence*, vol. 134, no. 1-2, pp. 181–199, 2002.

pdf.

References VI

[19] N. Papahristou, *Decision making in multiplayer environments: application in backgammon variants*.

PhD thesis, University of Macedonia, Department of Applied Informatics, 2015.

[20] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, pp. 354–359, 2017.

[21] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*.

Cambridge University Press, 2014.

440 pp.

References VII

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*.

MIT press, 2016.

http://www.deeplearningbook.org, 781 p.

[23] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*.

MIT press, 2018.

xvi + 486 pp.

[24] G. Rebala, A. Ravi, and S. Churiwala, *An Introduction to Machine Learning*.

Springer, 2019.

xxii + 263 p.

References VIII

[25] E. Alpaydin, *Introduction to machine learning (fourth edition)*.
Adaptive computation and machine learning, MIT press, 2020.
xxiv + 682 pp. 1st edition: 2004; 2nd, 2010; 3rd, 2014.

[26] F. Bach, "Learning theory from first principles, draft *(13-Oct-22)*," 2022.
https://www.di.ens.fr/~fbach/ltfp_book.pdf. xiv+360 p.

[27] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks*.
Cambridge University Press, 2022.
Draft: https://arxiv.org/pdf/2106.10165.pdf. viii+441 p.

[28] J. C. Ye, *Geometry of Deep Learning: A Signal Processing Perspective*.
No. 37 in Mathematics in Industry, Springer, 2022.

References IX

[29] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*.

Morgan kaufmann, 1988.

[30] D. C. Knill and W. Richards (editors), *Perception as Bayesian inference*.

Cambridge University Press, 1996.

Paper #1 is *Pattern theory: a unifying perspective*, by D. Mumford.

[31] J. Pearl *et al.*, *Causality: Models, reasoning and inference (2nd edition)*, vol. 19.

Cambridge University Press, 2009.

The first edition appeared in 2000.

References X

[32] M. Glymour, J. Pearl, and N. P. Jewell, *Causal inference in statistics: A primer*.

John Wiley & Sons, 2016.

[33] J. Pearl and D. Mackenzie, *The book of why: The new science of cause and effect*.

Basic Books, 2018.

[34] C. Bielza and P. Larrañaga, *Data-Driven Computational Neuroscience: Machine Learning and Statistical Models*.

Cambridge University Press, 2020.

xiv+747 p.

References XI

[35] S. Theodoridis, *Machine learning: a Bayesian and optimization perspective (2nd edition)*.

Academic Press, 2020.

xxvii+1131 p.

[36] B. Efron and T. Hastie, *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, vol. 6.

Cambridge University Press, 2021.

[37] S. Russell and P. Norvig, *Artificial intelligence: A modern approach (Fourth edition)*.

Pearson, 2022.

1166 p. Contributing writers: Ming-Wei Chang Jacob Devlin Anca Dragan David Forsyth Ian Goodfellow Jitendra M. Malik Vikash Mansinghka Judea Pearl Michael Wooldridge.

References XII

[38] K. P. Murphy, *Probabilistic Machine Learning: An introduction*.
MIT Press, 2022.

[39] E. T. Jaynes, *Probability theory: The logic of science*.
Cambridge University Press, 2003.

[40] J. Guttag, *Introduction to computation and programming using Python: With application to understanding data*.
MIT Press, 2016.
xvii + 447 p.

[41] A. Nandy and M. Biswas, *Reinforcement Learning: With Open AI, TensorFlow and Keras Using Python*.
Apress, 2018.
xiii + 167 p.

References XIII

[42] A. Said and V. Torra, *Data Science in Practice*.

Springer, 2019.

viii + 195 p.

[43] C. Lavor, S. Xambó-Descamps, and I. Zaplana, *A Geometric Algebra Invitation to Space-Time Physics, Robotics and Molecular Geometry*.

SBMA/Springerbrief, Springer, 2018.

[44] V. E. Balas, S. S. Roy, D. Sharma, and P. Samui, *Handbook of deep learning applications*, vol. 136.

Springer, 2019.

vi + 383 p.

References XIV

[45] F. Chollet, *Deep learning with Python (2nd edition, V4)*.

Manning Early Access Program, 2020.

210 p.

[46] A. Colomé and C. Torras, *Reinforcement Learning of Bimanual Robot Skills*, vol. 134 of *Springer Tracts in Advanced Robotics*.

Springer, 2020.

xx + 182 p.

[47] S. Xambó-Descamps (ed), *Systems, Patterns and Data Engineering with Geometric Calculi*.

ICIAM2019 SEMA SIMAI Springer Series, Springer, 2021.

Springer link.

References XV

[48] T. H. Davenport and S. M. Miller, *Working with AI: Real Stories of Human-Machine Collaboration*.

MIT Press, 2022.

[49] A. Rényi, "On measures of entropy and information," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 547–561, Berkeley, California, USA, 1961.

[50] J. Aczél and Z. Daróczy, *On measures of information and their characterizations*, vol. 168 of *Mathematics in science and engineering*.

Academic Press, 1975.

[51] N. Sajid, P. J. Ball, T. Parr, and K. J. Friston, "Active inference: demystified and compared," *Neural computation*, vol. 33, no. 3, pp. 674–712, 2021.

arXiv pdf.

References XVI

[52] L. Da Costa, P. Lanillos, N. Sajid, K. Friston, and S. Khan, "How active inference could help revolutionise robotics," *Entropy*, vol. 24, no. 3, p. 361, 2022.

[53] A. Barp, L. Da Costa, G. França, K. Friston, M. Girolami, M. I. Jordan, and G. A. Pavliotis, "Geometric methods for sampling, optimisation, inference and adaptive agents," 2022.

https://arxiv.org/pdf/2203.10592.