University of Brasilia

Institute of Exact Sciences
Department of Computer Science

THE INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

FREDERICO GUTH

Document presented for examination of
the Master Degree in Computer Science.

Supervisor
Prof. Dr. Teófilo Emidio de Campos

June 19, 2020

[version: June 19, 2020 at 18:52]

University of Brasilia

Institute of Exact Sciences
Department of Computer Science

THE INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

FREDERICO GUTH

Document presented for examination of
the Master Degree in Computer Science.

Prof. Teófilo Emidio de Campos (Supervisor)
CiC/UnB

Prof. John Shawe-Taylor
University College London

Prof. Moacir Antonelli Ponti
Universidade de São Paulo

Prof. Genaína Nunes Rodrigues
Computer Science Graduate Program Coordinator

Brasilia, June 19, 2020

[version: June 19, 2020 at 18:52]

ABSTRACT

In the last decade, we have witnessed a myriad of astonishing successes in Deep Learning. Despite those many successes, we may again be climbing a peak of inflated expectations. If in the past, the false solution was to throw computation power at problems, today we try trowing data. Such behaviour has triggered a winner-takes-all rush for data among a handful of large corporations, raising concerns about privacy and concentration of power. Yet, we know for a fact that learning from way fewer samples is possible: humans show a much better generalisation ability than our current state of the art artificial intelligence. To achieve such needed generalisation power, we must understand better how learning happens in deep neural networks. The practice of modern machine learning has outpaced its theoretical development, deep learning models present generalisation capabilities unpredicted by current machine learning theory. There is yet no established new general theory of learning which handles this problem. In 2015, Naftali Tishby and Noga Zaslavsky published a seminal theory of learning based on the information-theoretical concept of the bottleneck principle. This document aims to investigate the scattered efforts of using the information bottleneck principle to explain the generalisation capabilities of deep neural networks and consolidate them into a comprehensive digest of this new general deep learning theory.

RESUMO

Na última década, assistimos estupefatos uma miríade de sucessos em Aprendizado Profundo. Apesar de tamanho sucesso, talvez estejamos subindo um pico de expectativas infladas. No passado, incorremos no erro de tentar resolver problemas com maior poder computacional, hoje estamos fazendo o mesmo tentando usar cada vez mais

dados. Tal comportamento desencadeou uma corrida por dados entre grandes corporações, suscitando preocupações sobre privacidade e concentração de poder. Entretanto, é fato que aprender com muito menos dados é possível: humanos demonstram uma abilidade de generalização muito superior ao estado-da-arte atual em Inteligência Artificial. Para atingir tal capacidade, precisamos entender melhor como o aprendizado ocorre em Aprendizado Profundo. A pratica tem se desenvolvido mais rapidamente que a teoria na área. Modelos apresentam poder de genearlização que a atual teoria não explica. Em 2015, Naftali Tishby e Noga Zaslavsky publicaram uma teoria de aprendizado baseado no princípio do gargalo de informação (information bottleneck). Este documento visa investigar esforços esparços do uso do princípio do gargalo para explicar a capacidade de generalização de redes neurais profundas e consolidar tal conhecimento em um compêndio abrangente deste novo desenvolvimento teórico.

CONTENTS

ACRONYMS

AEP Asymptotic Equipartition Property

AI Artificial Intelligence

ANN Artificial Neural Network

CV Computer Vision

DL Deep Learning

DNN Deep Neural Network

GPU Graphical Processor Unit

IBT Information Bottleneck Theory

IT Information Theory

KB Knowledge Base

MLP Multilayer Perceptron

MLT Machine Learning Theory

NLP Natural Language Processing

SLT Statistical Learning Theory

SGD Stochastic Gradient Descent

NOTATION

This section provides a concise reference describing notation used throughout this document.

NUMBERS AND ARRAYS

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{a}^\frown\boldsymbol{b}$	vector $\boldsymbol{a}$ concatenated with vector $\boldsymbol{b}$
$\boldsymbol{A}$	A matrix
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns

SETS

A	A set
$\wp(A)$	The powerset (the set of subsets) of A
$\mathbb{X}, \mathbb{R}, \mathbb{N}, \cdots$	Special sets (or Spaces)
$\{0, 1\}$	The set containing 0 and 1
$\{0, \ldots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$a \in \mathbb{A}$	a is a member of the set $\mathbb{A}$
$\mathbb{B} \subset \mathbb{A}$	$\mathbb{B}$ is a subset of the set $\mathbb{A}$
$\mathbb{A} \cap \mathbb{B}$	The intersection of $\mathbb{A}$ and $\mathbb{B}$
$\mathbb{A} \cup \mathbb{B}$	The union of $\mathbb{A}$ and $\mathbb{B}$
$\overline{\mathbb{A}}$	The complement of $\mathbb{A}$
$\lvert\mathbb{A}\rvert$	The cardinality of $\mathbb{A}$

INDEXING

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
$\boldsymbol{A}_{i,j}$	Element ij of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$

LINEAR ALGEBRA OPERATIONS

$\boldsymbol{A}^\top$	Transpose of matrix $\boldsymbol{A}$
$\det(\boldsymbol{A})$	Determinant of $\boldsymbol{A}$

CALCULUS

$\frac{dy}{dx}$	Derivative of y with respect to x
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_S f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set S

Ω	A experiment or sample space
ω	An outcome (an example)
A	An event
$A \perp B$	The events A and B are independent
$X \perp Y$	The random variables X and Y are independent
$P(A \mid B)$	The probability of an event A given the event B happened
$P(X = a_i) \equiv P_X \equiv p(a_i) \equiv p_i \equiv p$	A probability distribution over a random variable (discrete or continuous defined by the context)
$a \sim p$	An example a drawn from distribution p
$\mathbb{E}_{x \sim p}[f(x)] \equiv \mathbb{E}_p f(x)$	Expectation of $f(x)$ with respect to $p(x)$
$\sigma^2(f(x))$	Variance of $f(x)$ under $p(x)$
$\mathrm{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $p(x)$
$H[X]$	Shannon entropy of the random variable X
$D_{KL}(p \Vert q)$	Kullback-Leibler divergence of distribution p and q
$\mathcal{N}(\boldsymbol{x}; \mu, \sigma^2)$	Gaussian distribution over $\boldsymbol{x}$ with mean μ and variance σ^2

FUNCTIONS

$f : \mathbb{A} \to \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta}) \equiv f_{\boldsymbol{\theta}}(\boldsymbol{x})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$
$\log_b x$	The logarithm base b of x
$\log x = \log_2 x$	If no base is specified, the base 2 is assumed
$\sigma(x)$	A non-linear activation function
x^+	Positive part of x , i.e., $\max(0, x)$
$\mathbb{1}_{[\text{condition}]}$	is the **indicator function** and is 1 if the condition is true, 0 otherwise

DATASETS AND DISTRIBUTIONS

We use the word **example** for an outcome drawn from a distribution and the word **sample** for a set of such *examples*. A dataset is a *sample*.

p_{data}	The data generating distribution
$\hat{p}_{\text{data}}$	The empirical distribution defined by the training set
$\mathbb{X}$	A set of training examples
x_i	The i-th example (input) from a dataset
$\boldsymbol{W}^{(i)}$	The matrix $\boldsymbol{W}$ of weights in the i-th layer of a network
y_i	The target associated with $\boldsymbol{x}_i$ for supervised learning
$\boldsymbol{X}$	The $m \times n$ matrix with input example $\boldsymbol{x}_i$ in row $\boldsymbol{X}_{i,:}$

1 INTRODUCTION

As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.

– Albert Einstein

This chapter contextualises the problem, define the objective and give an outline of this document.

1.1 CONTEXT

1.1.1 A Tale of Babylonians and Greeks

Figure 1.1: Richard Feynman, Nobel laureate physicist.

Richard Feynman (figure 1.1) used to lecture this story [17]: Babylonians were pioneers in mathematics; Yet, the Greeks took the credit. The Greek way of doing math is what we are used to: start from the most basic axioms and build up a system of knowledge from it. Babylonians were quite the opposite; they were pragmatic. No knowledge was considered more fundamental than others, and there was no urge to derive proofs in a particular order. Babylonians were concerned with the phenomena, Greeks with the ordinance. In Feynman's view, science is constructed in the Babylonian way. There is no fundamental truth. Theories try to connect dots from different pieces of knowledge. Only when science has advanced, one can worry about reformulation, simplification and ordering. Scientists are Babylonians; mathematicians are Greeks.

Mathematics and science are both tools for knowledge acquisition. Also, they are social constructs, as both rely on peer-reviewing. They are quite different, however.

Science is empiric, based on facts collected from experience. When physicists around the world measured events that corroborated Newton's *"Law of Universal Gravitation"*, they did not prove it correct; they just made his theory more and more plausible. Still, it was needed only one experiment to show that Einstein's *Relativity Theory* was even more believable. In contrast, we can and do prove things in mathematics.

In mathematics, knowledge is absolute truth, and the way one builds new knowledge with it, its inference method, is deduction. In science, knowledge is justified belief, there are degrees of plausibility, and its inference method is induction.

Mathematics is language, a formal one, a tool to precisely communicate some kinds of thoughts. As it happens with natural languages, there is beauty in it. The mathematician expands the boundaries of expression in this language and Greeks were poets. Even though Babylonians were first in finding mathematical truths, the Greeks invented Mathematics as epistemology.

Understanding the epistemic contrast between mathematics and science will help us understand the past of Artificial Intelligence (AI), and avoid some perils in its future. This contrast will be a recurrent theme in this dissertation.

1.1.2 The importance of theoretical development

In science, we collect facts, but they need interpretation. Science is a narrative of how we understand the world. A description without explanation is not science because it does not provide a plausible meaning to what we observed. This meaning brings with itself a view of how the world works, which can be applied in new situations and predict what will happen. It can be falsifiable.

To illustrate, take the ancient human desire of flying. Since antiquity, there have been stories of men strapping wings to themselves and attempting to fly by jumping from a tower and flapping those wings like birds (see figure 1.2). While the issues of lift, stability, and control were poorly understood, most attempts ended in severe injury or even death. It did not matter how much evidence, how many hours of seing different animals flying did those ludricous brave men experienced, the meaning they took from what they saw was wrong and their predictions incorrect.

They did not die in vain; Science advances when scientists are wrong. Theories must be falsifiable, and scientists cheer for their failure. When it fails, there is room for new approaches. Only when we understood the evidence of animal flight on the perspective of aerodynamics we learned to fly better than any other animal before. Science works by a

Figure 1.2: "A way of flying", Francisco Goya, 1815–1820, Amsterdam, Rijksmuseum.

"natural selection" of ideas, where only the fittest ones survive until a new better one is born. Those "researchers" deserved a Darwin award of science.

The Darwin Awards are satirical honors that recognize individuals who have unwillfully contributed to human evolution by selecting themselves out of the gene pool.

Being a Babylonian enterprise, science has no clear path. One of the interesting facts one can learn studying the history of science is that most powerful discoveries have arisen through the study of phenomena in man-made devices [44]. For instance, steam engines came before Carnot's work. These devices present a simplified small instance (easier to understand) of a more complex phenomena in nature. Another example is Information Theory. Several insights of Shannon's theory of communication were generalisations of ideas that were already present in telegraphy. A new general theory of artificial intelligence can for sure be developed from insights in the study of deep learning phenomena.

1.1.3 Bringing science to Computer Science

Despite the name, Computer Science has been more mathematics than science. We, computer scientists, are very comfortable with theorems and proofs; not much with theories.

Nevertheless, Artificial Intelligence (AI) has essentially become a Babylonian enterprise, a scientific endeavour. Thus, there is no surprise when some computer scientists still see AI with some distrust and even disdain:

- Even among AI researchers, there is a trend of "mathiness" and speculation disguised as explanations in conference papers [33].
- There is no place for papers that unpretentiously describe surprising phenomena without trying to come up with an explanation. As if the mere inconsistency of the current theoretical framework was unworthy of publication.

While physicists rejoice on finding phenomena that contradict current theories, computer scientists get baffled. In Natural Sciences, unexplained phenomena lead to theoretical development. In AI, they bring "winters".

1.2 PROBLEM

1.2.1 Learning Theory has failed deep

Herbert Simon (1916-2001) received the Turing Award in 1975, and the Nobel Prize in Economics in 1978

Artificial Intelligence has been through several "winters", periods of progress stagnation and lack of funding. In 1957, Herbert Simon famously predicted that within ten years, a computer would be a chess champion [45, section 1.3]. It took around 40 years, in any case. Computer scientists lacked understanding of the exponential nature of the problems they were trying to solve: Computational Complexity Theory had yet to be invented.

Machine Learning Theory (computational and statistical) tries to avoid a similar trap by analysing and classifying learning problems according to the number of samples required to learn them (and the number steps also). An honest assessment concludes it is now failing its mission. If by one hand, it lead to the development of useful machine learning algorithms like SVMs, by the other, it has also predicted that generalisation requires simpler models in terms of parameters, delaying the development of Deep Learning for years. In total disregard to the theory, deep learning models have shown spectacular generalisation power with hundreds of millions of parameters.

> *The curse of dimensionality is quite meaningless in practice[...][24].*
>
> *– Jeremy Howard,* FAST.AI *creator*

In the last decade, we have witnessed a myriad of astonishing successes in Deep Learning. Despite those many successes in research and industry applications, we may again be climbing a peak of inflated expectations. If in the past, the false solution was to throw computation power on problems, today we try trowing data. Such behaviour has triggered a winner-takes-all competition among a handful of large

corporations for who owns more data (our data), raising concerns about privacy and concentration of power.

Yet, we know for a fact that learning from way fewer samples is possible: humans show a much better generalisation ability than our current state of the art artificial intelligence. To achieve such needed generalisation power, we may need to understand better how learning happens in deep learning. Rethinking generalisation [67] will reshape the very foundations of machine learning theory.

1.2.2 Problem statement

The practice of modern machine learning has outpaced its theoretical development. In particular, deep learning models present generalisation capabilities unpredicted by current machine learning theory. There is yet no established new general theory of learning which handles this problem.

In 2015, Naftali Tishby and Noga Zaslavsky published a theory of learning based on the information-theoretical concept of the bottleneck principle [60]. This theory is general and can explain several deep learning phenomena inconsistent to current Machine Learning Theory. The reason it is still not yet *hors concours* is three-fold:

1. There has been some valid criticism to the experimental setting of the article mentioned above, which independent developments from Achille and Soatto address.
2. The understanding of this new theory demands a prior knowledge of Information Theory which deep learning practitioners of today are not used to.
3. Efforts on this new theory are scattered and knowledge still needs to be consolidated.

1.3 OBJECTIVE

This document aims to investigate the scattered efforts of using the information bottleneck principle to explain the generalisation capabilities of deep neural networks and consolidate them into a comprehensive digest of this new general deep learning theory.

1.4 OUTLINE

- Chapter 2 - Artificial Intelligence: The chapter defines what artificial intelligence is, presents the epistemological differences of intelligent agents in history, and discusses their consequences to the theory of machine learning.

- Chapter 3 - Probability Theory: The chapter derives propositional calculus and probability theory from a list of desired characteristics for skeptical agents.
- Chapter 4 - Information Theory: The chapter derives Shannon Information from Probability Theory, explicitises some implicit assumptions, and explains basic Information Theory concepts.
- Chapter 5 - Machine Learning Theory: The chapter presents the theoretical framework of Machine Learning, the PAC model, theoretical guarantees for generalisation, and expose criticism due to its lack of explanation on Deep Learning phenomena.
- Chapter 6 - Proposal: The chapter presents the plan for finishing the dissertation.

2 ARTIFICIAL INTELLIGENCE

I visualize a time when we will be to robots
what dogs are to humans,
and I'm rooting for the machines.

– Claude Shannon

This chapter defines what artificial intelligence is, presents the epistemological differences of intelligent agents in history, and discusses their consequences to the theory of machine learning.

2.1 ARTIFICIAL INTELLIGENCE

Definition 1. ***Artificial Intelligence*** *is the branch of Computer Science that studies general principles of intelligent agents and how to construct them [45].*

This definition uses the terms *intelligence* and *intelligent agents*, so let us start from them.

2.1.1 What is intelligence?

Despite a long history of research, there is still no consensual definition of intelligence[1]. Whatever it is, though, humans are particularly proud of it. We even call our species *homo sapiens*, as if intelligence was an intrinsic human characteristic.

In this document:

1 For a list with 70 definitions of intelligence, see [31].

Definition 2. ***Intelligence*** *is the ability to predict a course of action to achieve success in specific goals.*

2.1.2 Intelligent Agents

An agent is anything that perceives its environment and acts on it.

Under our definition, intelligence is not limited to humans. It applies to any agent: animal or machine. A bacteria can perceive its environment through chemical signals, process them, and then produce chemicals to signal other bacteria. An air-conditioning can observe changes of temperature, know its state, and adapt its functioning, turning off if it is cold or on if it is hot. *Intelligence exempts understanding*. The air-conditioning does not comprehend what it is doing. The same way as a calculator does not know arithmetics.

2.1.3 A strange inversion of reasoning

This is what the philosopher Daniel Dennett calls *Turing's strange inversion of reasoning*. The a idea of a *strange inversion* comes from one of Darwin's 19th century critics (MacKenzie [35] as cited by Dennett [15]):

> *In the theory with which we have to deal, Absolute Ignorance is the artificer; so that we may enunciate as the fundamental principle of the whole system, that,* ***in order to make a perfect and beautiful machine, it is not requisite to know how to make it****. This proposition will be found, on careful examination, to express, in condensed form, the essential purport of the [Evolution] Theory, and to express in a few words all Mr. Darwin's meaning; who, by* ***a strange inversion of reasoning****, seems to think Absolute Ignorance fully qualified to take the place of Absolute Wisdom in all of the achievements of creative skill.*
>
> – Robert MacKenzie

As counter-intuitive it was for MacKenzie [35] and many others to this date, intelligence can emerge from absolute ignorance. Turing's strange inversion of reasoning comes from the realization that his automata can perform calculations by symbol manipulation, proving that it is possible to build agents that behave intelligently, even if they are entirely ignorant of the meaning of what they are doing. They are useful ignorants.

2.2 DREAMING OF ROBOTS

2.2.1 From mythology to Logic

The idea of creating an intelligent agent is pehaps as old as humans. There are accounts of what we call artificial intelligence in almost any ancient mythology you look for: Greek, Etruscan, Egyptian, Hindu, Chinese [37]. For example, in Greek mythology, there is the story of the bronze automaton of Talos built by Hephaestus, the god of invention and blacksmithing, first mentioned around 700 BC.

This interest may explain why, since ancient times, philosophers have looked for *mechanical* methods of reasoning. Chinese, Indian and Greek philosophers all developed formal deduction in the first millennium BC. In special, Aristotelian syllogism, *laws of thought*, provided patterns for argument structures to yield irrefutable conclusions, given correct premises. This was the beginning of the field we call *Logic*.

2.2.2 Rationalism: The Cartesian view of the world

In the 13th century, the catalan philosopher Ramon Lull wanted to produce all statements the human mind could think of. For this task he developed "logic paper machines", discs of paper filled with esoteric colored diagrams that connected symbols representing statements.

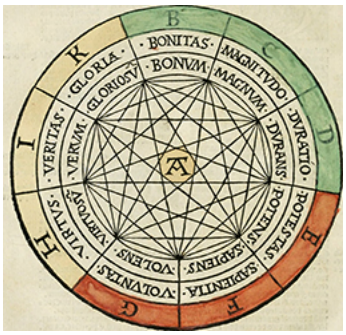


Figure 2.1: Example of one of Lull's *Ars Magna*'s paper discs.

In a modern reassessment of his work, unfortunately, *"it is impossible, pehaps, to avoid a strong sense of anticlimax"* [18]. His dilusional sense of importance, with a megolomaniac self-steem that hints psychosis, is more characteristic of cult founders. On the bright side, his ideas and books exerted some magic appeal that helped them to be rapidly disseminated through all Europe [18].

Lull's work greatly influenced Leibniz and Descartes, who, in the 17thcentury, believed that all rational thought could be mechanized. This was the basis of **rationalism**, the epistemic view of the *Enlightment*, that regarded reason as the sole source of knowledge. In other words, they believed that reality has a logical structure and that certain truths are *self-evidence*, and all truths can be derived from them.

During this period, there was a huge interest in developing artificial languages which later became what we now call formal languages.

> *If controversies were to arise, there would be no more need of disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down to their*

slates, and to say to each other: ***Let us calculate.***

– Gottfried Leibniz

The rationalist view of the world has had an enduring impact in society until today. In the 19th century, George Boole and others developed a precise notation for statements about all kinds of objects in the world and the relations among them. Before them, logic was philosophical rather than mathematical. The name of Boole's masterpiece, *"The Laws of Thought"*, is a great indicative of his Cartesian worldview.

In the beginning of the 20th century, some of the most famous mathematicians, David Hilbert, Bertrand Russel, Alfred Whitehead, were still interested in formalism: they wanted mathematics to be formulated on a solid and complete logical foundation. In special, Hilbert's *Entscheidungs Problem* (decision problem) asked if there were limits to the power of mechanical logic proofs.

Kurt Gödel's incompletness theorem (1931) basically proved that any language expressive enough to describe arithmethics of the natural numbers is either incomplete or inconsistent. This is a limit to logic systems. There will always be truths that will not be provable from within such languages: i.e. there are true statements that are undecidable.

Alan Turing brought a new perspective to the *Entscheidungs Problem*: a function on natural numbers that cannot be represented by an algorithm in a formal language cannot be computable. Gödel's limit appears in this context as functions that are not computable, e.g. there is no algorithm that can decide whether another algorithm will stop or not (the halting problem). To prove that, Turing developed a whole new general theory of computation: what is computable and how to compute it; laying out a blueprint to build computers, and making possible the research of Artificial Intelligence as we know it. An area in which Turing himself was very much invested.

2.2.3 Empiricism: The skeptical view of the world

The response to **rationalism** was **empiricism**, the epistemological view that knowledge comes from sensory experience, our perceptions of the world. As Locke put it *"there is nothing in the intellect that was not previously in the senses"*. Bacon, Locke and Hume were great exponents of this movement, which established the grounds of the scientific method.

David Hume, in particular, presented in 18th century a radical empiricist view: reason only does not lead to knowledge. In [27], Hume distinguishes *relations of ideas*, propositions which derive from deduction, and *matters of facts*, which rely on the connection of cause and effect through experience, induction. Hume's critiques, known

as *the Problem of Induction*, added a new slant on the debate of the scientific method

From Hume's own words:

Figure 2.2: David Hume.

> *The bread, which I formerly eat, nourished me; that is, a body of such sensible qualities was, at that time, endued with such secret powers: but does it follow, that other bread must also nourish me at another time, and that like sensible qualities must always be attended with like secret powers? The consequence seems no wise necessary.*
>
> – David Hume

We cannot apply deduction and expect the future will resemble the past, since suggesting that it will not does not contradict logic. Yet, we do expect uniformity in Nature. As we see more examples of something happening, more evidence, it is *wise* to expect that it will happen in the future just as it did in the past. There is no rationality in this expectation, though.

Hume explains that we see conjunction repeatedly, e. g. "bread" and "nourish", and we expect *uniformity in nature*, we expect that "nourish" will always follow "eating bread"; When we fulfill this expectancy, we misinterpret it as causation. In other words, we project causation into phenomena. Hume explicitized that this connection does not exist in Nature. We do not "see causation," we create it.

This projection is *Hume's strange inversion of reasoning* [26]: We do not like sugar because it is sweet; Sweetness exists because we like (or need) it. There is no sweetness in honey. We wire our brain so that glucose triggers a labelled desire we call sweetness. As we will see later, sweetness is *information*. This insight shows the pattern matching nature of humans. Musicians have relied on this for centuries. Music is a sequence of sounds in which we expect a pattern. The expectancy is the tension we feel while the chords progress. When the progression finally *resolves*, forming a pattern, we release the tension. We feel pattern matching in our core. It is very human, it can be very helpful and wise, but it is irrational.

The epistemology of the skeptical view of the world is science: to weight ones beliefs to the evidence. Knowledge is not absolute truth, but justified belief. It is a Babylonian epistemology.

As we will demonstrate in this document, in rationalism, what connects knowledge and good actions is logic. In empiricism, the connection between knowledge and justifiable actions is determined by probability. More especifically, by the Bayes' theorem. As Jaynes puts it, probability theory is the logic of science [28].

The Bayes' theorem is attributed to the reverend Thomas Bayes, after the posthumously publication of his work. By the publication time, it was an already known theorem, derived by Laplace.

2.2.4 The birth of AI as a research field

In 1943, McCulloch and Pitts, a neurophysiologist and a logician, demonstrated that neuron-like eletronic units that act and interact by physiologically plausible principles could be wired together and perform complex logical calculation [45]. Moreover, they showed that any computable function could be computed by some network of connected neurons [38]. This was the birth of Artificial Neural Networks (ANNs), even before the field of AI had this name. It was also the birth of connectionism, the approach of using artificial neural networks, loosely inspired by biology, to explain mental phenomena and imitate intelligence. Their work was an inspiration to John von Neumann's demonstration on how a universal Turing machine can be created out of electronic components, which lead to the advent of computers and programming languages. Ironically, these advents hastened the ascent of the formal logicist approach called symbolism, in disregard to connectionism.

In 1956, John McCarthy, Claude Shannon (who invented Information Theory, figure 2.3), Marvin Minsky and Nathaniel Rochester organized a 2-month summer workshop in the Darthmouth College with the goal of bringing researchers of different fields concerned with "thinking machines"(cybernetics, information theory, automata theory). The workshop attendees became a community of researchers and term *"artificial intelligence"* was chosen for the field.

Figure 2.3: Claude Shannon.

2.3 BUILDING INTELLIGENT AGENTS

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind

–John Godfrey Saxe, The Blind Men and the Elephant [47]

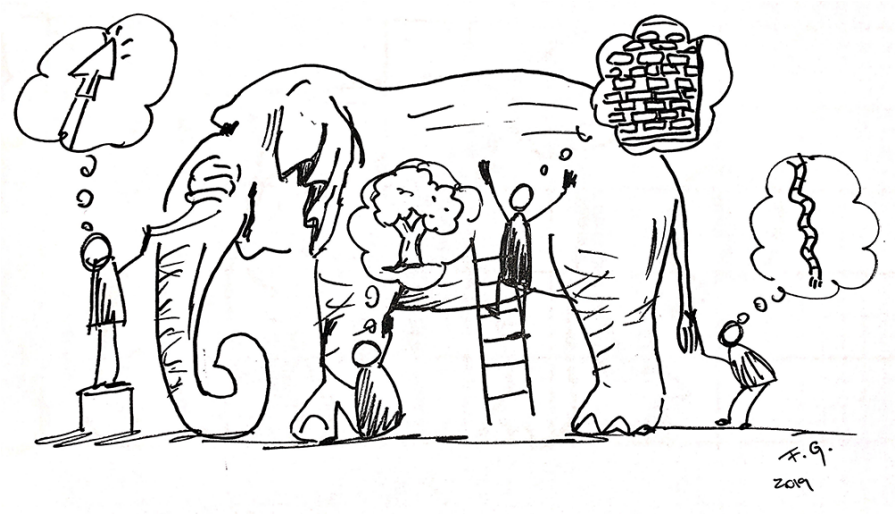


Figure 2.4: The Blind Men and the Elephant.

2.3.1 Anatomy of intelligent agents

Like in the parable of "the blind men and the elephant", an intelligent agent shall model her understanding of the world[2] from limited sensory data.

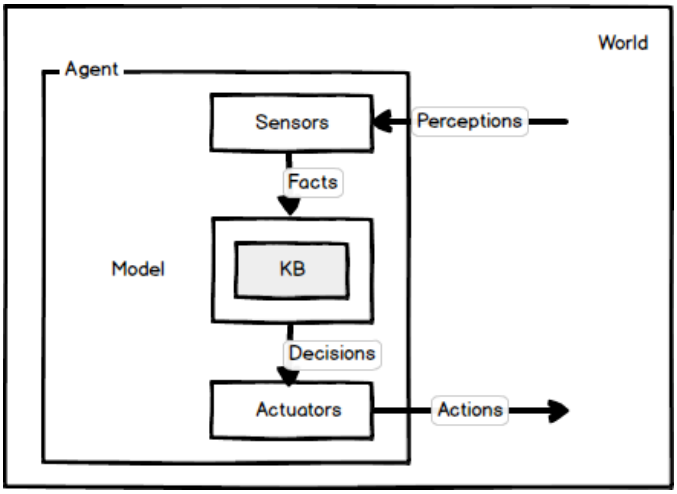


Figure 2.5: Anatomy of an Intelligent Agent

Thus, an agent perceives her world with sensors, treat sensory data as facts and use these facts to possibly update its model of the world, use the model to decide her actions, and acts via her actuators. In a way, agents are continually communicating with the world in a perception/action conversation (figure 2.5).

The expected result of this conversation is a change in the agent's Knowledge Base (KB), therefore in her model and, more importantly, her future decisions. The model is an abstraction of how the agent

2 We are using the word *world* as in [56], to denote *environment* or *Nature*.

"thinks" the world is (her "mental picture" of the environment). Therefore, it should be consistent with it: if something is true in the world, it is equally true, *mutatis mutandis*, in the model. A Model should also be as simple as it can be such that the agent can make decisions that maximize a chosen performance measure, but not simpler. As the agent knows more about the world, less it gets surprised by it.

This rudimentary anatomy is flexible enough to entail different epistemic views, like the rationalist (mathematical) and the empirical (scientific); different approaches to how to implement the knowledge base (it can be learned, therefore updatable, or it can be set in stone from expert prior knowledge); and also from how to implement it (a robot or a software).

Noteworthy, though, is that the model that transforms input data into decisions should be the target of our focus.

2.3.2 Symbolism

Symbolism is the pinneacle of rationalism. In the words of Thomas Hobbes, one of the forerunners of rationalism, *"thinking is the manipulation of symbols and reasoning is computation"*. Symbolism is the approach to building intelligent agents that does just that. It attempts to represent knowledge with a formal language, and explicitly connects the knowledge with actions. It is *competence from comprehension*. In other words, it is *programmed*.

Even though McCulloch and Pitts work on artificial neural networks predates Von Neumann's computers and the increase of interest in the study of "thinking machines" of the 1950s that would become AI, symbolism dominated the field up until the 1980s. It was so ubiquitous that nowadays, symbolic AI is even called "good old fashioned AI".

The symbolic approach can be traced back to Nichomachean Ethics [9]:

> *We deliberate not about ends but about means. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade, nor a statesman whether he shall produce law and order, nor does anyone else deliberate about his end. They assume the end and consider how and by what means it is to be attained; and if it seems to be produced by several means they consider by which it is most easily and best produced, while if it is achieved by one only they consider how it will be achieved by this and by what means this will be achieved, till they come to the first cause, which in the order of discovery is last.*
>
> – Aristotle

This perspective is so entrenched that Russell, Norvig, and Davis [45, p. 7] still says: *"only by understanding how actions can be justified can we understand how to build an agent whose actions are justifiable"*; even though, in the same book, they cover machine learning (which we will address later in this chapter) without noticing it is a proof that

there are other ways to build intelligent agents. Moreover, it is also a negation of competence without comprehension. The reality is that even for many AI researchers, the strange inversion of reasoning is uncomfortable. All humans, even those in prisons and under mental health care, think their actions are justifiable. Is that not an indication that we rationalise our actions post factum?

2.3.2.1 *Claude Shannon's Theseus*

After writing what is probably the most important master's dissertation of the 20th century and "inventing" the Information Theory that made possible the Information Age we live today, Claude Shannon enjoyed the freedom to pursue any interest to which his curious mind led him [55]. In the 1950s, his interest shifted to building artificial intelligence. He was not a typical academic, in any case. A lifelong tinkerer, he liked to "think" with his hand as much as with his mind. Besides developing an algorithm to play chess (when he even did not have a computer to run it), one of his greatest achievements in AI was Theseus, a robotic maze-solving mouse.

Many AI students will recognise in Theseus the inspiration to Russell, Norvig, and Davis [45]'s Wumpus world.

To be more accurate, Theseus was just a bar magnet covered with a sculpted wooden mouse with copper whiskers; the maze was the "brain" that solved itself.

> *"Under the maze, an electromagnet mounted on a motor-powered carriage can move north, south, east, and west; as it moves, so does Theseus. Each time its copper whiskers touch one of the metal walls and complete the electric circuit, two things happen. First, the corresponding relay circuit's switch flips from "on" to "off," recording that space as having a wall on that side. Then Theseus rotates 90° clockwise and moves forward. In this way, it systematically moves through the maze until it reaches the target, recording the exits and walls for each square it passes through.*
>
> *– Klein [30]".*

2.3.2.2 *Symbolic AI problems*

Several symbolic AI projects sought to hard-code knowledge about domains in formal languages, but it has always been a costly and slow process that could not be scaled.

Anyhow, by 1965 there were already programs that could solve any solvable problem described in logical notation [45, p.4]. Hubris and lack of philosophical perspective, however, made computer scientists believe that "intelligence was a problem about to be solved[3]."

Those inflated expectations lead to disillusionment and funding cuts. They failed to estimate the inherent difficulty in slating informal knowledge in formal terms: the world has many shades of grey. Besides, complexity theory had yet to be developed, and they did not

3 Marvin Minsky, head of the artificial intelligence laboratory at MIT (1967)

count on the exponential explosion of the problems they were dealing with.

2.3.3 Connectionism: a different approach

Connectionism is an approach to the study of human cognition that utilizes mathematical models, known as connectionist networks or artificial neural networks. It was pioneered by McCulloch and Pitts in 1943 [38]. One of the most famous developments of the first wave of connectionism was Frank Rosenblatt's Perceptron, an algorithm for learning binary classifiers, or more specifically threshold functions:

$$y = \begin{cases} 1 \text{ if } \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} > 0 \\ 0 \text{ otherwise} \end{cases} \tag{2.1}$$

where $\boldsymbol{W}$ is the vector of weights, $\boldsymbol{x}$ is the input vector, $\boldsymbol{b}$ is a bias and $\boldsymbol{y}$ is the classification. In the context of neural networks, a perceptron is an artificial neuron using a step function as the activation function.

The fundamental idea in connectionism is that **intelligent behaviour emerges from a large number of simple computational units when networked together** [19].

See figure 2.6, termites self-cooling mounds keep the tempture inside at exactly 31°C, ideal for their fungus farming; while the temperatures outside range from 2 to 40°C through the day. Such building techniques inspired architect Mike Pearce to design a shopping mall in Zimbabwe that uses a tenth of the energy used by a conventional building of the same size.

Where does termites intelligence come from?

> *Individual termites react rather than think, but at a group level they exhibit a kind of cognition and awareness of their surroundings. Similarly, in the brain, individual neurons don't think, but thinking arises in the connections between them.*
>
> — Radhika Nagpal, Harvard University [36].

Such collective intelligence happens in groups of just a couple of million termites. In the human brain there are around 80 to 90 billion neurons, each one of them less capable than a termite, but collectively they show still uncomparable intelligence capabilities.

In contrast with the symbolic approach, in neural networks the knowledge is not explicit in symbols, but implicit in the strengh of the connections between the neurons. Besides, it is a very general and flexible approach, since these connections can be updated algorithmically: they are algorithms that *learn*: the connectionist approach is an example of what we now call Machine Learning.

Throughout AI history, there has been a pendulum between connectionism and symbolism. Connectionism has been through several

(a) Building in Harare, Zimbabwe is modelled after termite mounds. Photo by Mike Pearce.

(b) Cathedral termite mound, Australia. Photo by Awoisoak Kaosiowa, 2008.

Figure 2.6: Biomimicry of termite technique achieves superior energy efficiency in buildings.

"winters", due to theoretical developments that pointed to some flaws of a current practice. Such "winters" delayed AI for several years. As an example, backpropagation was derived around 1962 by Stuart Dreyfus. Still, due to the lack of interest in connectionism, his work was ignored until after Rumelhart, Hinton and Williams *reinvented* it in 1986.

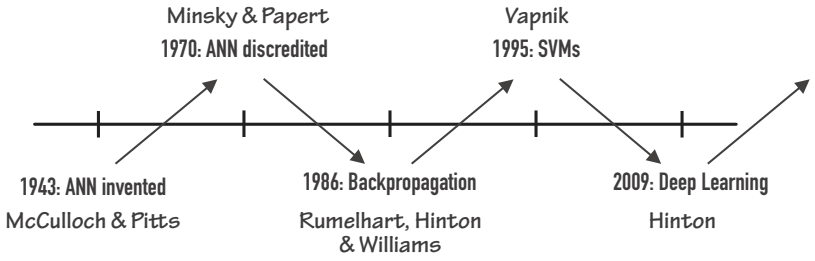


Figure 2.7: A brief history of connectionism.

2.3.4 Machine Learning

Look at the figure 2.8. Is this a picture of a cat? How would you write a program to do such simple classification task (cat/no cat)? You could come up with clever ways to use *features* from the input picture and process them to make a guess. Clearly, though, it is not an easy program to design. Worse, even if you manage to program such task, how much would it worth to accomplish a related task, to recognise a dog, for example?

For long, this was the problem of researchers in many areas of interest of AI: Computer Vision (CV), Natural Language Processing

Figure 2.8: Is this a picture of a cat?

(NLP), Speech Recognition; a lot mental effort was put, with very poor results, in problems that we humans solve with ease.

The solution is a completly different approach for building artificial intelligence: instead of building the program to do the task, build the program that outputs the program that does the task. In other words, learning algorithms use what we call "training data" to infere the transformations to the input that generates the desired output. It is competence without comprehension.

2.3.4.1 *Types of learning*

Machine Learning can happen in different scenarios, which differ in the availability of training data, how training data is received, and how the test data is used to evaluate the learning. Here, we describe the most typical of them [43]:

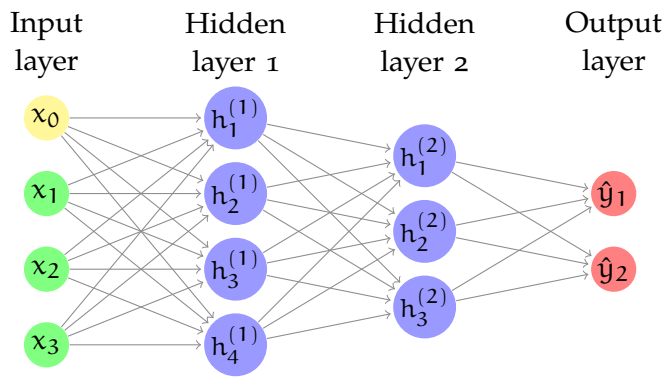
- **Supervised learning:** The most successful scenario; the learner receives a set of labelled examples as training data and makes predictions for unseen data.
- **Unsupervised learning:** The learner receives unlabeled training data, and makes predictions for unseen instances.
- **Semi-supervised learning:** The learner receives a training sample consisting of both labelled and unlabelled data, and makes predictions for unseen examples. Semi-supervised learning is usual in settings where unlabeled data is easily accessible, but labelling is too costly.
- **Reinforcement learning:** The learner actively interacts with the environment, and receives an immediate reward for her actions. The training and testing phases are intermixed.

2.3.5 Deep Learning

The 2010_s have been an AI Renaissance not only in academia but in the industry as well. Almost all the successes are due to Deep Learning (DL), in particular, supervised deep learning with vast amounts of data trained in Graphical Processor Units (GPUs). It was the decade of DL.

"Deep learning algorithms seek to exploit the unknown structure in the input distribution in order to discover good representations, often at multiple levels, with higher-level learned features defined in terms of lower-level features Bengio [11]".The name is explained by one of his students: *"A graph showing the concepts being built on top of each other is a deep graph. Therefore the name, deep learning [19]"*. Although it is a direct descendent of the connectionist movement, in its modern form it goes beyond the neuroscientific perspective. It is more a general principle of learning multiple levels of compositions.

The quintessential example of a deep learning model is the feedforward deep network or Multilayer Perceptron (MLP) [45].



Definition 3. *Let,*

$\boldsymbol{x}$ *be the input vector* $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m\}$

k *be the layer index, such that* $k \in [1, l]$,

d_k *is the number of elements in the hidden layer* (k),

$\boldsymbol{W}_{i,j}^{(k)}$ *be the matrix of weights in the* k*-th layer, where* $i \in [0, d_{k-1}], j \in [1, d_k]$ *and* $\boldsymbol{W}_{0,:}^{(k)}$ *are the biases*

σ *be a non-linear function,*

a ***Multilayer Perceptrons (MLPs)*** *is a neural network where the input is defined by:*

$$h^{(0)} = 1^\frown \boldsymbol{x}, \tag{2.2}$$

a hidden layer is defined by:

$$h^{(k)} = \sigma^{(k)}(\boldsymbol{W}^{(k)\top} h^{(k-1)}). \tag{2.3}$$

The output is defined by:

$$\hat{y} = h^{(l)}. \tag{2.4}$$

Deep Learning is usually associated with Deep Neural Networks (DNNs), but the network architecture is only one of its components:

1. DNN architecture
2. Stochastic Gradient Descent (SGD) – the optimiser
3. Dataset
4. Loss function

The architecture is not the sole component essential to current Deep Learning success. The SGD plays a crucial role, and so does the usage of large datasets.

A known problem, though, is that DNNs are prone to overfitting[4]. Zhang et al. [67] shows that state-of-the-art convolutional deep neural networks can easily fit a random labelling of training data.

4 Overfitting is explained in section 5.3.

3 PROBABILITY THEORY

A wise man proportions his belief to the evidence.

– David Hume

In this chapter, propositional calculus and probability theory are derived from a list of desired characteristics for skeptical agents.

3.1 FROM LANGUAGE TO PROBABILITY

3.1.1 Formal Languages

We, as intelligent agents, do not know how the world is; we only know how we perceive it. Our ideas are mental pictures of how we imagine it. Like the blind men and the elephant (section 2.3), how do we know, then that our model is the same as someone else's? *Communicating*. We need to communicate with each other to check if our mental picture of the world, our model, is consistent with the experience of others[1].

We use language to describe the world. However, natural languages, like English, German, Portuguese, are ambiguous, and we need contextual clues and other information to more clearly communicate meaning. To avoid this, an intelligent agent uses a formal language.

A formal language is a mathematical tool created for precise communication about a specific subject. For example, arithmetic is a language for calculations. Chemists have a language that represents chemical structures of molecules. Programming languages are formal languages that express computations. In a nutshell, a formal language is a set of

1 We can take this idea further and think that at any moment we need to communicate with our past selfs to check if new evidence is consistent with our prior model.

words (strings) whose letters (symbols) are taken from an alphabet and are well-formed according to a specific set of rules, a grammar:

$$\text{Let } L =< \Sigma, \Phi >, \text{ be a formal language where:} \tag{3.1}$$
$$\Sigma = \{S_1, S_2, \cdots, S_n\} \text{ is an alphabet,} \tag{3.2}$$
$$\Phi = \Phi_1 \cup \Phi_2 \cup \cdots \cup \Phi_k \text{ is a set of operations,} \tag{3.3}$$
$$\tag{3.4}$$

where:

Φ_1 is the set of unary operations,

Φ_2 is the set of binary operations,

...

Φ_k is the set of k-ary operations.

A formal language allows a quantitative description of a state of knowledge and defines how this state can be updated on the presence of new evidence.

With this definition, we can also think that a formal language is what Sowinski [56] calls a *realm of discourse*, i. e. all the valid formed *strings*[2] that one can derive; everything one can *say* about the world.

Interestingly, formal languages allow us to manipulate representations of the world without dealing with their semantics. They are the basis of *"Turing's strange inversion,"* (see section 2.1.3) by doing allowed operations on strings, computers can compute in a superhuman speed and accuracy without ever comprehending what they are doing.

3.1.2 From Rationalism to Propositional Calculus

RATIONAL AGENTS Rational agents are agents that can form representations of a complex world, use deduction as the process of inference to derive updated representations, and use these new representations to decide what to do. In other words, rational agents are the consequence of the epistemological view of *rationalism*.

When a particular statement's truth value is established by a rational agent, all statements that are formed in her knowledge base from that statement instantly feel that update. A rational agent cannot hold contradictions.

DESIDERATA FOR A RATIONAL LANGUAGE We want to build a language for rational agents with the following desired characteristics:

I. **knowledge is absolute**; a sentence[3] can be either true or false;

2 Strings, words, sentences, propositions, formulae are names used interchangeably through the literature.

3 A sentence can be either a single symbol or a string formed with several symbols according to the grammar.

II. **unambiguous**, a constructed sentence can only have one meaning;

III. **consistent**; a language without paradoxes, i. e. whatever path chosen to derive a sentence truth value will lead to the same assignment;

IV. **minimal**; uses the smallest set of symbols possible.

Let $L_R =< \Sigma_R, \Phi_R >$ be the formal language built from these constraints; where sentences are either axiom symbols or compounded sentences formed using special symbols called operators, each operator denoting one operation $\phi \in \Phi_R$.

It is possible to prove that L_R only needs one operator [28, 56]: NAND (or XOR), and it is also equivalent to Propositional Calculus. In other words, Logic is the language that emerges from our desiderata, from rationalism. **Logic is the language of mathematics**.

Proposition is synonym to sentence, and Propositional Calculus is also known as Sentential Calculus.

A point worth mentioning is that using Logic as an agent formal language means the **implicit acceptance** of the constraints above.

3.1.3 From Empiricism to Probability Theory

The constraints that lead to Logic are very restrictive to use in the real world; the rational language has a comparatively small realm of discourse. Hume would say that it is only useful for *relations of ideas*, talking in the abstract, and not for *matters of facts*, talking about reality.

A realm of discourse to talk about reality needs the empiricist perspective where knowledge is justified belief, and that one should *weights her beliefs to the evidence.* The quantity that specifies to what degree we believe a proposition is true is constrained by other beliefs, i.e. by previous experience and evidence gathering.

SKEPTICAL AGENTS In the skeptical agent[4], the agent derived from empiricist epistemology, beliefs are not independent of each other [13], they form an interconnected web that is the agent's knowledge base. The update mechanism, its inference method, follows the principle of minimality, i.e. it tries to minimise the change in the knowledge base.

DESIDERATA FOR A SKEPTICAL LANGUAGE As we did for rational agents, let us state a set of desired characteristics for the language of science, $L_S =< \Sigma_S, \Phi_S >$[5]:

I. **Knowledge is a set of beliefs, quantifiable by real numbers and dependent on prior evidence:**

4 Other authors have called these agents epistemic agents [13], idealised epistemic agents [56] or robots [28].

5 [13, 28, 56] also present this same idea of deriving probability theory from a desiderata.

Let $S_i \in \Sigma_S$ be sentences about the world. Given any two statements S_1, S_2, the agent must be able to say that S_1 is more plausible than S_2, or that S_2 is more plausible than S_1, or that S_1 and S_2 are equally plausible [13]. Thus we can list statements in an increasing plausibility order. Real numbers can represent this transitive ordering [13][6].

Using $(S|K)$ in a function is a notation abuse that we accept in order to better explain the idea.

Let b be a measure of degrees of belief in S given K:

$$b : \Sigma_S \to \mathbb{R} \tag{3.5}$$

$$b : S \mapsto b(S|K) \tag{3.6}$$

Here we capture the fact that plausibility (degrees of belief) is not a function of a sentence, but a relation between a sentence and a given assumed prior knowledge.

II. **Common sense:**
Plausibility of compound sentences should be related to the plausibility of the sentences that form them.
We already showed that a minimal rational language has only one operator. Here, instead of using the NAND operator, for a matter of familiarity, let us use the almost minimal language with the operators NOT ($\neg$) and AND ($\wedge$). In this setting, we are saying $\exists\ f,\ g$ such that [56]:

(a) $b(\neg S|K) = f[b(S|K)]$

(b) $b(S_1 \wedge S_2|K) = g[b(S_1|K), b(S_1|S_2), b(S_2|K), b(S_2|S_1)]$

III. **Consistency:**
The functions f and g must be consistent with the grammar Φ (production rules).
Consistency guarantees that whatever path used to compute the plausibility of a statement in the context of the same knowledge web (the same set of constraints) must lead to the same degree of belief.

(a) Beliefs that depend on multiple propositions cannot depend on the order in which they are presented.

(b) No proposition can be arbitrarily ignored.

(c) Propositions which are known to be identical must be assigned the same degree of belief.

Such desiderata have a name; it is known as Cox's axioms, and one can derive the Sum Rule and the Product Rule (see section 3.4) from them, therefore, also the Bayes' Theorem (section 3.9), and reverse engineer Kolmogorov's Axioms of Probability Theory (section 4) [13, 28, 56, 58].

6 We are implicitly assuming that the language we are building has infinite statements. A further discussion on this continuity assumption can be found in [56, p. 26]

In other words, Probability Theory is the language that emerges from our desiderata, from empiricism. **Probability theory is the logic of science** [28], and our measure b is what is usually called probability P.

Again, here we explicit the fact that by using Bayesian inference to build and communicate concepts of the world (models), we are assuming the Cox's axioms above.

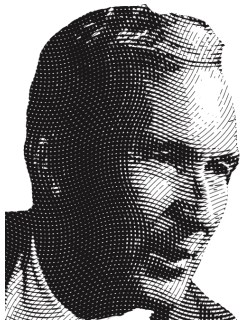


Figure 3.1: Andrey Kolmogorov.

3.1.4 Assumptions and their consequences

As a side note, let us take this opportunity to explore what some assumptions mean to human intelligence in particular. It is indisputable[7] that humans are not rational, neither skeptical agents. The whole idea of imagining a skeptical agent is a consequence of wanting to address intelligence without the complexities of humans.

However, are humans irrational because of biology or psychology? Are we irrational for lack of will or could it be that Nature wires the human brain in a way that pr*events* us from following these axioms? Here we argue that biology has an important role. Researchers have found, for instance, that visual acuity can be permanently impaired if there is a sensory deficit during early post-natal development [64]. If the human brain is not exposed to some samples in its infancy, it will never achieve the accuracy level it could achieve if it had experienced them, regardless of experiencing those examples later. In other words, *human beliefs depend on the order in which pieces of evidence are presented*, contradicting Cox's axiom III.(a).

3.2 FORMALIZING PROBABILITY THEORY

Up to this point in the chapter, our purpose was to give an intuition of what probability is. We did that by showing it is possible to derive the Kolmogorov's Axioms (section 4) from the empiricist epistemology. In this section, we will use these axioms to formalise concepts in Probability Theory and provide visual explanations whenever possible to the jargon commonly used in this field.

Worth of mention, several concepts in this section are *relations of ideas*, not *matters of fact*. The probability of an *event* E, P(E), can be computed by marginalisation (see subsection 3.8), but as discussed before, there are no beliefs in a vacuum. In reality, there is only the probability of an *event* E given some background knowledge K. This change of epistemological perspective is essential to be remembered

7 Unless you are an economist.

now; that we will expose the idealised development of concepts in Probability Theory.

3.3 EXPERIMENTS, SAMPLE SPACES AND EVENTS

The set of possible outcomes of an ***experiment*** is the **sample space** Ω. Let us use the canonical *experiment* of rolling a dice. In this experiment, the sample space is:

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$$

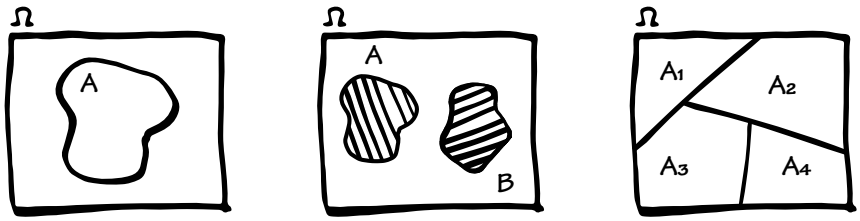
An **outcome** or **realization** is a point $\omega \in \Omega$:

$$\omega_3 = \boxdot$$
$$\Omega = \{\omega_1 = \boxdot, \cdots, \omega_6 = \boxdot\}.$$

An **Event** is something you can say about the *experiment*, e. g. "The dice rolled to the an odd number". It is a true proposition. But easier than writing so much, we denote *events* with letters. **Events are subsets of** Ω (see figure 3.2a).

$$A = \{a_1 = \boxdot, a_2 = \boxdot, a_3 = \boxdot\}$$
$$A \subset \Omega$$

We say that $A_1, A_2, \cdots$ are **mutually exclusive** or **disjoint** *events* if $A_i \cap A_i = \emptyset, \forall i \neq j$. For example, A is the *event* "the dice rolled to the value 5" and B is the *event* "the dice rolled to an even number". In this case, A and B are disjoint (see figure 3.2b)



(a) An *event* A.

(b) Disjoint events A and B: $A \cap B = \emptyset$.

(c) A partition of Ω: $\bigcup_i A_i = \Omega$.

Figure 3.2: Events, disjoint events and partitions.

A **partition** of Ω is a sequence of disjoint sets A_i (see figure 3.2c), where:

$$A_1, A_2, \cdots A_i \text{ s.t. } (A_1 \cup A_2 \cup A_3 \cdots = \bigcup_{i=1}^{\infty} A_i) = \Omega \tag{3.7}$$

3.4 PROBABILITY

The powerset of Ω, $\wp(\Omega)$, is the set of all possible subsets of Ω.

Definition 4 (Kolmogorov's Axioms)**.** *A function* $P : \wp(\Omega) \to \mathbb{R}$ *that maps any* event A *to a real number* $P(A)$ *is called the* ***probability measure*** *or a* ***probability distribution*** *if it satisfies the Kolmogorov's axioms [63]:*

Axiom 1. $P(A) \geqslant 0, \forall A$

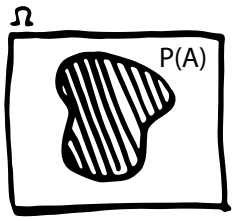
Axiom 2. $P(\Omega) = 1$

Axiom 3. *If* A *and* B *are disjoint, i.e.* $A \perp B$,

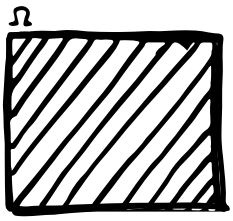
$$P(A \vee B) = P(A) + P(B) \tag{3.8}$$

(Sum Rule)

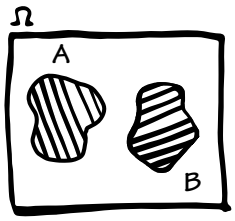
Visually, we can represent the probability of an *event* A, $P(A)$, as the proportion of the sample space, the *event* occupies. To diferentiate *events* from their probabilities, we will shade the area of the *event*.



(a) Axiom 1: $P(A) \geqslant 0$

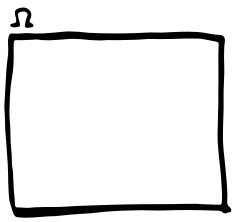


(b) Axiom 2: $P(\Omega) = 1$.

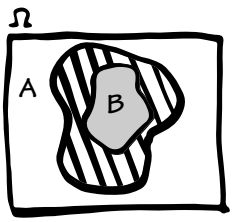


(c) Axiom 3: $A \cap B = \emptyset \implies P(A \vee B) = P(A) + P(B)$.

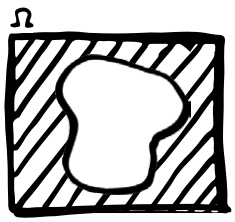
Figure 3.3: Kolmogorov's Axioms



(a) $P(\emptyset) = 0$.



(b) $A \subset B \to P(A) \leqslant P(B)$.



(c) $P(\bar{A}) = 1 - P(A)$.

Figure 3.4: Direct derivations from Kolmogorov's axioms.

Directly from the Kolmogorov Axioms, one can derive [28] other properties (see figure 3.4):

$$P(\emptyset) = 0 \tag{3.9}$$
$$B \subset A \implies P(B) \leqslant P(A) \tag{3.10}$$
$$0 \leqslant P(A) \leqslant 1 \tag{3.11}$$
$$P(\bar{A}) = 1 - P(A). \tag{3.12}$$

3.5 JOINT EVENT

Definition 5. *A joint* event A, B *is the set of outcomes where:*

$$(A, B) = \omega \in \Omega : (\omega \in A \cap B)$$

Therefore,

$$P(A, B) = P(\omega \in \Omega : (\omega \in A \cap B))$$

Notation hell:

$P(A, B)$
$\equiv P(B, A)$
$\equiv P(A \wedge B)$
$\equiv P(A \cap B)$
$\equiv P(A \times B)$.

When talking about *events* as propositions, it is straightforward to use logic notation $P(A \wedge B)$, but when we start to use *random variables* (section 3.10), we will adopt the shorthand notation $P(A, B)$.

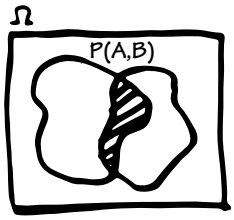


Figure 3.5: Joint event (A, B)

3.6 INDEPENDENT EVENTS

Definition 6. *Events* A *and* B *are* ***independent*** *(*$A \perp B$*) if:*

$$A \neq \emptyset, B \neq \emptyset \implies P(A) > 0, P(B) > 0 \tag{3.13}$$
$$P(A, B) = P(A \wedge B) = P(A) \cdot P(B) \tag{3.14}$$
(Product Rule)

Disjoint *events* cannot be independent, since (from equation 3.13) $P(A) \cdot P(B) > 0$, but as disjoint *events* (figure 3.2b) $P(A \wedge B) = P(\emptyset) = 0$, leading to contradiction.

Independence can be assumed or derived by verifying:

$$''P(A \wedge B) = P(A) \cdot P(B). \tag{3.15}$$

(Independent variables)

3.7 CONDITIONAL PROBABILITY

As we have explained before (section 3.1.3), the plausability of an outcome or a set of outcomes depends on a web of interconnected prior beliefs. So, in reality, as a *matter of fact*, what exists are probabilities *conditional* to a given prior assumption.

Definition 7. *If* $P(B) > 0$ *then the* ***conditional probability*** *of* A *given* B *is:*

$$P(A|B) \triangleq \frac{P(A,B)}{P(B)} \tag{3.16}$$

Also:

$$P(A,B) \triangleq P(A|B) \cdot P(B) \tag{3.17}$$

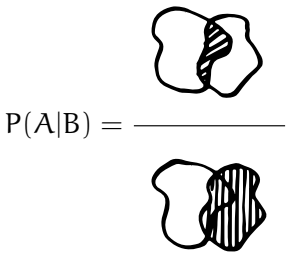


Figure 3.6: Conditional probability of A given B.

Except if $P(A) \equiv P(B)$, $P(A|B) \neq P(B|A)$. Also, $P(A|B) = P(A) \iff A \perp B$.

Venn diagrams are a not helpful to see that the events *are independent [63], as it all depends on the areas of intersection and the sizes of A and B, which are quite difficult to estimate without computational help.*

3.8 MARGINAL PROBABILITY

Theorem 1. *Let* $A_1, \cdots, A_k$ *be a partition of* Ω*. Then, for any* event B*,*

$$P(B) = \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i) \tag{3.18}$$

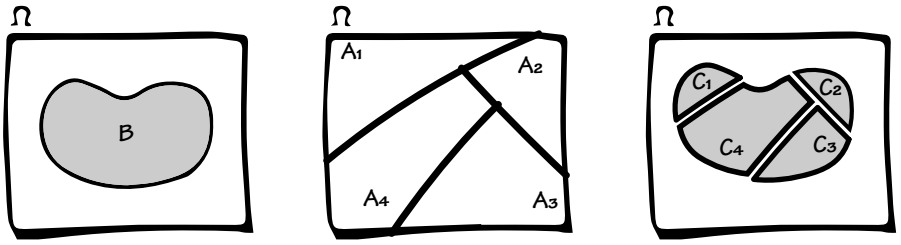


Figure 3.7: An *event* B, a partition A_i over Ω, and $C_i = (B, A_i)$.

Remember: $(B, A) \equiv (B \cap A)$.

Proof. Define $C_i = (B, A_i)$. Let $C_1, \cdots C_k$ be disjoint and $B = \bigcup_{i=1}^{k} C_i$. Therefore:

$$P(B) \triangleq P(\bigcup_{i=1}^{k} C_i) \tag{3.19}$$

$$\overset{3.8}{=} \sum_i P(C_i) \tag{3.20}$$

$$\triangleq \sum_i P(B, A_i) \tag{3.21}$$

$$\overset{3.16}{=} \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i) \qquad \text{(Law of Total Probability)}$$

□

3.9 BAYES' THEOREM

Theorem 2. *Let* $A_1, \cdots, A_k$ *be a partition of* Ω *s.t.* $P(A_i) > 0, \forall i$ *then,* $\forall i = 1, \cdots, k$*:*

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_i P(B|A_i) \cdot P(A_i)} \qquad \text{(Bayes' theorem)}$$

Proof. From equations 3.16, 3.17 and 3.18:

$$P(A_i|B) \overset{3.16}{=} \frac{P(A_i, B)}{P(B)} \tag{3.22}$$

$$\overset{3.17}{=} \frac{P(B|A_i) \cdot P(A_i)}{P(B)} \tag{3.23}$$

$$\overset{3.18}{=} \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} \tag{3.24}$$

□

We call $P(A_i)$ the **prior** of A, and $P(A_i|B)$ the **posterior** probability of A.

3.10 RANDOM VARIABLES

Definition 8. *A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω, $\omega \mapsto X(\omega)$.*

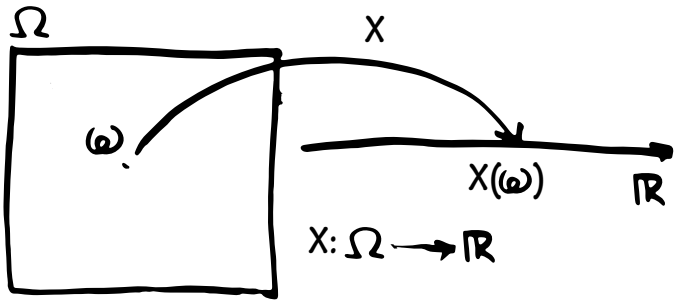


Figure 3.8: Random variable.

Given a random variable X, the probability of an outcome x can be expressed as:

$$P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) = x\}) \tag{3.25}$$

Several works on Probability Theory choose to start by defining random variables, rarely mentioning sample spaces, *events* or the connection with logical propositions.

This usual approach is, nevertheless, confusing. Beyond the fact that random variables are not variables, but functions; nor random, they model uncertain *events*; it is hard to grasp what random variables are without understanding their reasons for being.

3.10.1 Notation hell

If a *random variable* is a function, how can we write $P(X = 4)$ or $P(X > 7)$? The reason for such confusion is due to some notation abuse that became standard in works on probability theory. It is not easy to grasp it in the beginning, but the explanation was already stated at equation 3.25. $P(X = x)$ is a shorthand for $P(X^{-1}(x))$.

Although technically, a *random variable* is a function, in practice, it is just a mathematical tool to help us associate propositions with numbers. That said, it is called a *random variable* because the notation abuse treat the function as a variable.

To help clear up such confusion, let us recap a little the notation we have established before:

In the canonical *experiment* of rolling a dice, instead of writing the plausability of the proposition *"The dice will roll to number 4"* is $\frac{1}{6}$, it is easier to assign a letter to the proposition, or as we called the *event*. Let us use *event* D to represent the proposition. Then, we can use

$P(D) = \frac{1}{6}$. Now, we are going one step further, instead of using the *event* D we use the *random variable* D, in italic, and say $P(D = 4) = \frac{1}{6}$.

Notice the difference between a *random variable* and an *event*: D could assume any value (even $D = 7$ which is clearly outside of our *sample space*). Would it not be easier then to use an index to the *event* letters, i. e. D_4 to value 4, and D_1 to value 1, etc.? Not really.

An event *can be seen as a special kind of* random variable. *E. g. , a random variable D is the truth function (also known as the indicator function) over an event* D*:*

$$D = \mathbb{1}_D$$

*That is the reason one can say that "*random variables *define* events*."*

Besides providing this shorter notation, the mapping of the random variable allows us to manipulate *events* as numbers: for example, we can chart probability distributions using random variables, what we cannot cope with *events*.

3.11 PROBABILITY DISTRIBUTIONS

Definition 9. *A probability distribution of a discrete random variable X or* ***probability mass function (pmf)*** *is a function $p : \Omega \to [0, 1]$ that provides the probabilities of occurrence of different possible outcomes in an* experiment *(sample space):*

$$p(x) = P(X = x), \qquad \text{(pmf)}$$

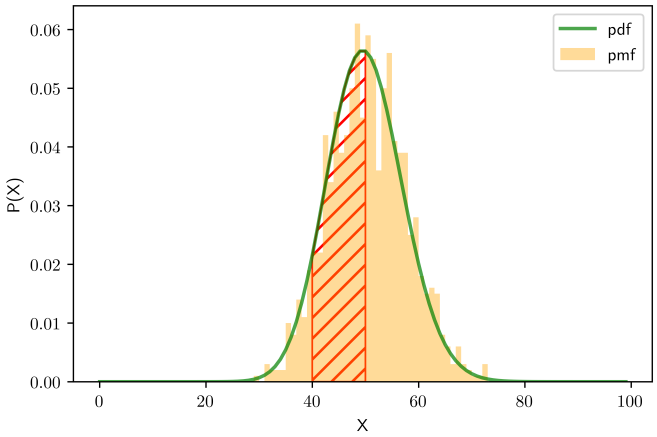


Figure 3.9: Probability mass function, probability density function, and probability of an interval (hatched area).

If X is countinuous, $P(X = x) \to 0$, therefore we need to use intervals in this case.

Definition 10. *A probability distribution of a countinous random variable X in an interval A, or* ***probability density function (pdf)*** *is a function $p(x)$ that measures the probability of randomly selecting a*

value within the interval $A = [a, b]$*, as the area under its curve for the interval A:*

$$P(A) = P[a \leqslant X \leqslant b] = \int_a^b p(x)\,dx, \text{ and:} \tag{3.26}$$

$$\begin{cases} p(x) \geqslant 0, \forall x \\ \int_{\mathbb{R}} p(x)\,dx = 1 \end{cases} \tag{pdf}$$

Now that we explained what distributions are, here we highlight some useful distributions:

3.11.1 Uniform distribution

$X \sim \text{Uniform}(a, b)$, if:

$$p(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & x \notin [a, b] \end{cases}$$

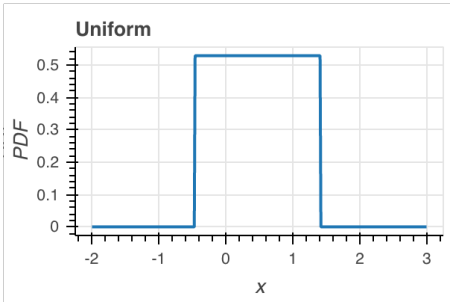


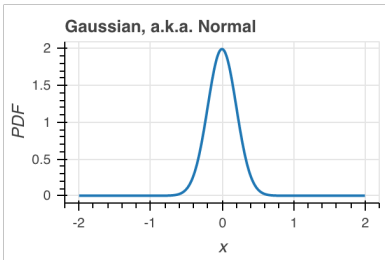
Figure 3.10: Uniform distribution

3.11.2 Normal distribution

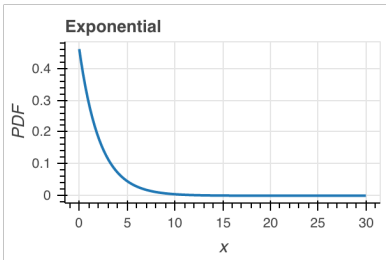
$X \sim \mathcal{N}(\mu, \sigma^2)$, if:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}$$

where $\mu \in \mathbb{R}$ (mean) and $\sigma > 0$ (standard deviation). We say that X has a **standard Normal distribution** if $\mu = 0$, $\sigma = 1$.



(a) Gaussian distribution, also known as the *normal*.



(b) Exponential distribution.

Figure 3.11: Gaussian and Exponential distributions.

3.11.3 Exponential distribution

$X \sim \text{Exp}(\lambda)$, if:

$$p(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0, \\ 0 & x < 0. \end{cases}$$

where $\lambda > 0$ is the *rate parameter* of the distribution.

3.12 JOINT DISTRIBUTIONS

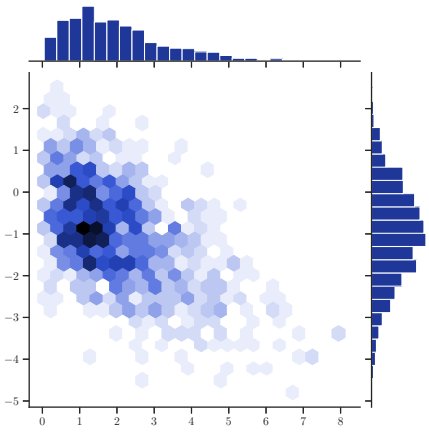


Figure 3.12: A chart of a joint distribution.

Definition 11. *Given a pair of discrete random variables X and Y, we define the **joint mass function** by $p(x, y) = P(X = x, Y = y)$.*

Definition 12. *Given a pair of continuous random variables X and Y, we define the* ***joint density function*** *by* $p(x, y)$*, where:*

i. $p(x, y) \geqslant 0$

ii. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y)\, dxdy = 1$

iii. $\forall A \subset \mathbb{R} \times \mathbb{R}, P((X, Y) \in A) = \int \int_A p(x, y)\, dxdy.$

3.13 EXPECTANCY, VARIANCE AND COVARIANCE

Definition 13. *The* ***expected value*** *or* ***mean*** *of X is:*

$$\mathbb{E}(X) = \langle X \rangle = \sum\!\!\!\!\!\!\int_x x\, p(x)\, dx = \mu = \mu_X \tag{3.27}$$

Theorem 3. *Let* $X_1, \cdots, X_n$ *be random variables and* $a_1, \cdots, a_n$ *be constants, then from the* Sum Rule*:*

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i(\mathbb{E}(X_i)) \tag{3.28}$$

Theorem 4. *Let* $X_1, \cdots, X_n$ *be independent random variables, then from the* Product Rule*:*

$$\mathbb{E}(\prod_i X_i) = \prod_i \mathbb{E}(X_i) \tag{3.29}$$

Definition 14. *Let X be a random variable with mean* μ*. The* ***variance*** *of X is defined by:*

$$\sigma^2 = \sigma_X^2 = \mathbb{E}(X - \mu)^2 \tag{3.30}$$

assumming this expectation exists. The standard deviation is σ*.*

Definition 15. *Let X and Y be random variables with means* μ_X *and* μ_Y*, and with standard deviations* σ_X *and* σ_Y*. The* ***covariance*** *between X and Y is defined as [63, p.74]:*

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \tag{3.31}$$

and the correlation as:

$$\rho = \rho_{X,Y} = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3.32}$$

Theorem 5. *The covariance satisfies:*

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{3.33}$$

3.14 INDEPENDENT SAMPLING

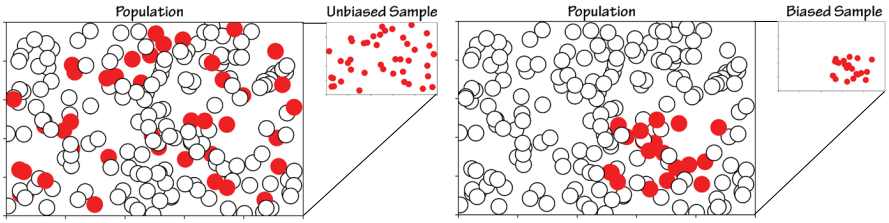


Figure 3.13: An i.i.d. sample (left) and a biased sample (right). Adapted from [41].

In this document, an element of a sampling is called an example. *A* sample *is a set of examples drawn from a distribution.*

One common assumption in Machine Learning Theory is that examples are *identically and independently distributed – i.i.d.* This means that the probability of obtaining a first training example (x_1, y_1) does not affect which (x_2, y_2) will be draw in the next observation.

This is useful wherever a census of the population of interest, knowing all possible values, is unfeasible. This is the usual case, data analysis is carried out using a sample to represent the population. When the sample is i.i.d., each example in the population has the same chance of being observed (figure 3.13 – left).

If there is a constraint on which examples of the population are actually sampled, we say that the sample is *biased* (figure 3.13 – right).

3.15 BIBLIOGRAPHICAL REMARKS

There are many important topics in Statistics that were left out from this short review chapter, where the focus was to present the most important concepts that will be used later on in this document.

Overall, the chapter was influenced by Wasserman [63]. The idea to derive Probability Theory from Logic can be found in Jaynes [28], Sowinski [56] and Caticha [13], being this last resource one of the most complete in the subject.

4 INFORMATION THEORY

Only through communication can human life hold meaning.

– Paulo Freire

This chapter derives Shannon Information from Probability Theory, explicitises some implicit assumptions, and explains basic Information Theory concepts.

4.1 FROM PROBABILITY TO INFORMATION

In section 2.3.1, we exposed that an agent updates its model of the world from sensory data, experience. We have also shown how this update happens; a skeptical agent *proportions her beliefs to the evidence* according to Bayes' theorem.

The amount of this update on knowledge is not uniform. Some experiences are more valuable than others, i.e. some evidence will produce a bigger change in the agent's knowledge, leading to a greater impact in her future actions. We say that those experiences are more informative.

Definition 16. ***Information*** *is what changes belief [13, 56].*

Let us say that an agent's *prior* belief in a statement S is $P(S)$. After experiencing some evidence e, her *posterior* set of beliefs is updated to incorporate the evidence, $P(S|e)$. The prior and the posterior are related by the product rule (section 3.6) [56]:

$$P(S|e) = \frac{P(e|s)}{P(e)} \cdot P(S) \tag{4.1}$$

$P(S)$ *is in fact* $P(S|K)$*, but we supress it to reduce the clutter.*
Here we are talking about events*:* $P(S|e)$ *is a short hand for* $P(S|\{e\} \wedge K)$.

We shall call this ratio by which prior and posterior are related as likelihood ($\mathcal{L}$):

$$\mathcal{L}(e;S) = \frac{P(S|e)}{P(S)} \tag{4.2}$$

$$P(S|e) = \mathcal{L}(e;S) \cdot P(S). \tag{4.3}$$

Simply by observing equation 4.2, we can conclude that if information (i) is what changes belief, information and likelihood must be related to one another:

$$i_S(e) = f(\mathcal{L}(e;S)). \tag{4.4}$$

Moreover, if an experience does not change a belief, it contains no information: $f(1) = 0$.

We also hope that when the likelihood changes by an infinitesimal amount the information does not change discontinuously, so f is continuous.

The information gathered from independent *events* must reflect the commutativity of Cox's axiom III. (a). Let $\mathcal{L}_1 = \mathcal{L}(e_1;S)$ and $\mathcal{L}_2 = \mathcal{L}(e_2;S)$, information must satisfy the functional constraints [56]:

$$\begin{cases} f(\mathcal{L}_1 \wedge \mathcal{L}_2) & = f(\mathcal{L}_1) + f(\mathcal{L}_2) \\ f(1) & = 0 \\ f & \text{is continuous.} \end{cases}$$

This functional form can be solved and its solution is [13]:

$$f = A \cdot \ln \mathcal{L}(e;S) \therefore$$
$$i_S(e) = A \cdot \ln \mathcal{L}(e;S). \tag{4.5}$$

From equations 4.5 and 4.2,

$$i_S(e) = A \cdot \ln \frac{p(S|e)}{p(S)} \tag{4.6}$$

$$i_S(e) = A \cdot \ln p(S|e) - A \cdot \ln p(S). \tag{4.7}$$

The constant A allow us to use any base b in the logarithm:

$$A = \frac{1}{\ln b} \rightarrow i_S(e) = \log_b p(S|e). \tag{4.8}$$

We can argue that the ammount of information gained by the agent about the world is equivalent to some ammount of *hidden information* h that was revealed to the agent by the *event* e.

Hence, $i_S(e) = -\Delta h(e)$, from eq. 4.7:

$$i_S(e) = \log p(S|e) - \log p(S) \tag{4.9}$$

$$i_S(e) = -\left[\left(\underbrace{-\log p(S|e)}_{h(S|e)}\right) - \left(\underbrace{-\log p(S)}_{h(S)}\right)\right] \tag{4.10}$$

$$i_S(e) = -\Delta h(e). \tag{4.11}$$

Delightfully, our definition of *hidden information* that reduces the uncertatinty of the agent, and emerged from our definition of information,

$$h(S) = -\log p(S) \tag{4.12}$$

is equivalent to Shannon's self information:

Also known as the Shannon information content of an outcome [34].

$$I[S] = -\log p(s) \tag{4.13}$$

In Information Theory (IT), self-information is defined as the entropy contribution of an individual message (or symbol); in other words, how much uncertainty reduction can be attained by an individual *event*. This uncertainty reduction is precisely what we derived.

Shannon's information can be derived from probability theory (c.q.d.).

4.2 EXPLICITING THE IMPLICIT ASSUMPTIONS

When we derived Probability Theory from a language for rational agents, and then Information Theory from Probability Theory, we exposed their assumptions. Including **consistency**:

i. A belief on a statement can not depend on the path used to arrive at it. In other words, it does not matter the order in which evidence is presented.

ii. No evidence can be arbitrarily ignored.

iii. Statements which are known to be identical must be assigned the same degree of belief.

By using these theories as the basis of Machine Learning Theory, we are **implicitly accepting** the assumptions above. Symbolic AI guarantees that their agents follow such assumptions by construction.

We know humans do not follow these assumptions and the whole point of conceptualizing rational agents was to study a simplified form of intelligence. For humans,

i. the order in which we experience pieces of evidence matter [64];

ii. we forget or suppress past experiences;

iii. we can change our mind even in the absence of new evidence.

What about Deep Neural Networks?

There is nothing by construction that forces DNNs to be consistent. Recent findings [3] show that, in DNNs, the order in which evidence is presented has a significant effect on the learning result. Therefore, we conjecture:

Conjecture 1. *A complete learning theory of Deep Learning (DL) has to address* ***time*** *and its effect on the* ***cost*** *of changing a belief.*

4.3 SHANNON'S MATHEMATICAL THEORY OF COMMUNICATION

Information Theory (IT) has an identifiable beginning: Shannon's 1948 paper "A mathematical theory of communication" was a giant leap towards understanding communication and defining *information*. Despite his acknowledging of the influence from previous works by pioneers such Harry Nyquist and Ralph Hartley, it was Shannon's unifying vision that revolutionized communication and provided a *"blueprint"* for the information age [8]. His theory defines unbreachable limits, the *"laws of information"* [57]:

In a rare piece of collaboration, Shannon asked his lunchroom table colleagues at Bell Labs to come up with a snapier name than binary digit. Bigit *was considered, but John Tukey's proposal,* bit, *was chosen [55].*

i. There is an upper limit, the **channel capacity**, to the amount of information that can be communicated through the channel;

ii. **Noise** reduces the **channel capacity**;

iii. There is an encoding that allows **lossless** communication trough a **noisy channel**.

The idea that it is possible to transmit information with zero error through a noisy channel is not at all intuitive and its theorethical proof was a totally unexpected result at the time. In the following sections, we will explain the concepts of IT that allow us to comprehend these *laws of information*.

4.3.1 The communication problem setting

Shannon deliberately chose not to deal with fuzzy concepts as intelligence or meaning:

> ***The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.*** *Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These* ***semantic aspects of communication are irrelevant to the engineering problem****. The significant aspect is that the actual message is one* ***selected from a set*** *of possible messages.*
>
> *– Claude Shannon, "A mathematical theory of communication", p.1*

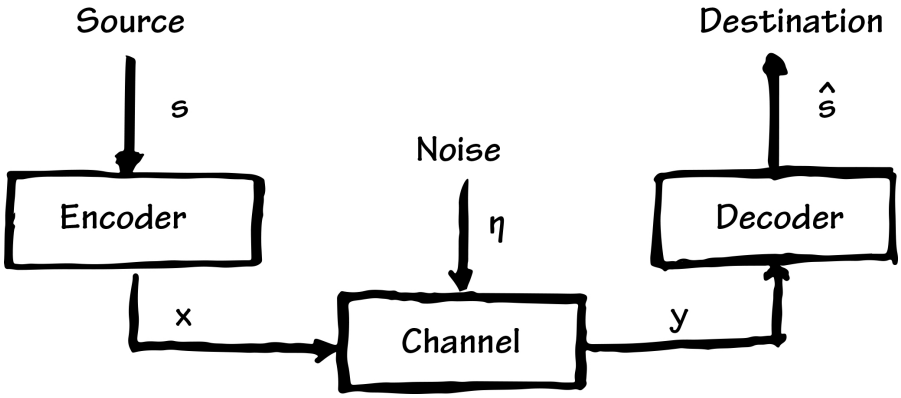


Figure 4.1: The communication problem setting.

Conceptually, this setting can be explained as follows (figure 4.1): The Source:

1. selects a message s from a set of possible messages $\mathbb{A}_S$,
2. encodes the message s into a string of symbols x, the signal,
3. transmits this string of inputs x through a noisy channel.

s is the intended message. One can think about it as the meaning or the semantics.

$\mathbb{A}_S$ is the alphabet or the set of possible outcomes of the random variable S.

The Destination, then:

1. receives a string of symbols y,
2. decodes the string y into the most probable message $\hat{s}$.

4.4 INFORMATION

The reason for communication is to change another agent's behaviour. In other words, *communication either affects the conduct of the recipient, or it is like it has never happened* [48, p.100]. We have already established (section 4.1, definition 16) that *information is what changes belief*; thus, changes an agent's conduct. So, **communication is transmitting information**.

Noteworthy, information is independent of the *enconding* or the channel chosen. You can use any language (English, Portuguese, music, images, dance, etc.) and any means of transmission (letter, telegraphy, microwaves, etc.) that the transmitted information remains the same.

To simplify, Shannon constrained semantics to the act of choosing a message from a set of finite possibilities. A source (a person, a machine or phenomenon) that always sends the same message never surprises the receiver, and the message carriers no information. On the contrary case, a source that sends symbols at random is impossible to predict and, therefore, every message carries maximal information.

Therefore, in this setting, *information is a measure of freedom of choice in selecting the message* [49, p.100]. In other words, it is a measure of surprisal, or of uncertainty reduction.

In the aforementioned famous paper, Shannon limited to say that mathematically, if the set of possible messages $\mathbb{A}_S$ is finite, any function of the size of this set $f(|\mathbb{A}_S|)$ is a measure of information, and that the logarithmic function is a natural choice. We shall expand on this idea.

4.4.1 A guessing game

Imagine a number from 1 to 1000. Let us assume that you picked the number at random. Thus, each number in the range had the same chance of being chosen, $\frac{1}{1000}$. How many questions do I need to ask to guess your number correctly? Well, it depends on what are the allowed answers. I could ask:

- How many hundreds has your number?
- Then, I would ask how many tens has your number?
- Then, how many units?

In this case, the number of questions needed is three, the height of the tree in figure 4.2. This is because we allowed each answer to be a *digit*, therefore the *branching factor* b of the decision tree was 10. It is easy to notice that the height of the tree is $\log_b(1000)$.

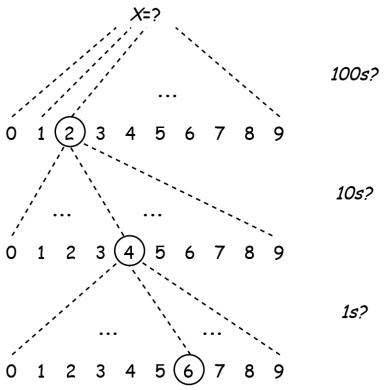


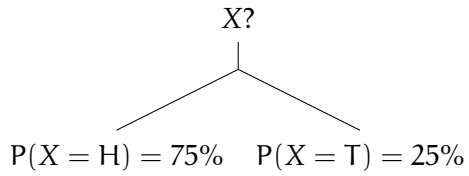
Figure 4.2: Branching factor of 10 to find 246.

It is now clear what Shannon meant by saying that the logarithmic function was the natural measure of information. The logarithm will give the height (number of questions) of the decision tree. The branching factor is just a measurement unit and can be chosen arbitrarily.

The smallest branching factor is 2, a *bit*. So, one bit is the amount of information resulted from choosing between two equally likely options.

To solve the same guessing game with *bits*, i. e. with yes or no questions, one proceeds with a binary search, and in the worse case it will need $\log_2(1000) = \frac{\log_{10}(1000)}{\log_{10}(2)} \approx 9.96 \therefore 10$ questions.

How about if the choice was among not equally likely options? Let us examine the simplest case of an unfair coin.



Here, we expect that the outcome to be *heads*, so if it turns *tails* we get surprised. Before the coin flip, we were 25% certain (our belief measure) that the *experiment* would turn *tails*. If it, in fact, turns *tails*, our certainty reaches 100%, growing by a factor of $\frac{1}{0.25} = 4$. So its reasonable to think that our uncertainty of the *tails* outcome decreasead by a factor of 4 as well. We were 75% certain that the *experiment* would turn *heads*. If it in fact turns *heads*, our uncertainty of the *heads* outcome decreasead by a factor of $\frac{1}{0.75} \approx 1.\overline{3}$. How do we transform this uncertainty reduction factor to a measure in bits. In other words, how do we measure in bits the information gained by unvealing an outcome?

Notice that 1 *bit* is the amount of information that reduces uncertainty from 2 possible states to 1, a factor of 2. Also, 2 bits of information reduce the uncertainty from the 4 possible representable states with 2 bits to 1, a factor of 4.

$$
\begin{aligned}
2^1 \text{ factor} &= 1 \text{ bit} \\
2^2 \text{ factor} &= 2 \text{ bits} \\
&\dots \\
2^n \text{ factor} &= n \text{ bits} \\
\therefore x \text{ factor} &= \log_2(x) \text{ bits}
\end{aligned}
$$

So, if an outcome has probability $p(x)$:

$$
\frac{1}{p(x)} \text{ factor} \implies \log_2 \frac{1}{p(x)} \text{bits} = -\log_2 p(x) \text{ bits}
$$

If the factor is a measure of the reduction in freedom of choice, the factor is the information gained by knowing the outcome of the *experiment*. This is known as **self-information** or information content of an outcome:

Definition 17. *The **information content, self-information, surprisal**, or **Shannon information** of a particular outcome* x *of an* experiment *is defined as:*

$$
I[x] = h[x] = -\log p(x) \tag{4.14}
$$

(information content of outcome)

As we already had derived in section 4.1.

4.4.2 Entropy

Information theory magnitudes are functions of the probabilities random variables and not directly of random variable. To address this difference, we opt to use square brackets instead of parenthesis.

In practice, however, we are not usually interested in the information of a particular outcome, but in how much surprise, on average, is expected for the entire set of possible outcomes.

Definition 18. *The entropy* $H[X]$ *of a random variable* X *is defined to be the average Shannon information content of its possible outcomes:*

$$H[X] \triangleq \mathbb{E}_p \frac{1}{\log p(x)} = -\sum_{x \in \mathbb{A}_X} p(x) \log p(x) \text{ bits/symbol}. \quad (4.15)$$

We will constrain our explanations of Information Theory to the discrete case. It can be argued that if we are interested in models that will be used by computers, some quantization will always happen.

Entropy can be seen in two ways:

1. as the quantity of information "produced" by the source [49, p.18].
2. as a measure of *uncertainty* or lack of pattern.

Average information shares de same definition as entropy, therefore, to known whether a quantity is information or entropy depends whether it is given or taken [57]. In other words, uncertainty reduced is information gained, and vice-versa. If a random variable X is very uncertain, then it has high entropy. If we are told the outcome of the variable $X = x_j$, we have been given information that is equal to the uncertainty we had. Thus, receiving an amount of information is equivalent to having the same amount of entropy taken away.

4.5 THE SOURCE

In the problem setting proposed by Shannon, the source generates a message, symbol by symbol. The choice of each symbol depends on the "preceding choices as well as the particular symbols in question" [49, p.10].

A mathematical model that follows this description is known as a *stochastic process*. Any discrete source can be represented by a stochastic process. "Conversely, any stochastic process may be considered a discrete source" [49].

Definition 19. *A* ***stochastic (or random) process*** *is a set of random variables indexed by a variable* $i \in \mathbb{N}$ *(usually representing time):*

$$S_i, i \in \mathbb{N} \quad (4.16)$$

(Stochastic Process)

In the original formulation, Shannon modeled the source as a stochastic process indexed by time. He thought the source as an entity that emits a certain rate, amount of information (bits) per period (seconds):

$$R_S \triangleq \frac{H[S]}{T_S} \frac{\text{bits}}{\text{second}}, \tag{4.17}$$

where T_{src} is the average time in seconds of transmitting a symbol. For simplification sake, from now on we will just say that the source rate is:

$$R_S = H[S] \text{ bits/symbol} \tag{4.18}$$

4.5.1 Markov chains

More specifically, Shannon proposed using a special kind of stochastic process called an *ergodic Markov chain* to model the source.

Definition 20. *An order-k **Markov chain** is a stochastic process that satisfies the following property:*

$$P(S_i|S_{i-1}, S_{i-2}, \cdots, S_{i-k}) = P(S_i|S_{i-1}, S_{i-2}, \cdots, S_1) \tag{4.19}$$

*The **ergodic** property means statistical homogeneity [49]: its statistical properties can be deduced from a single, sufficiently long, random sample of the process.*

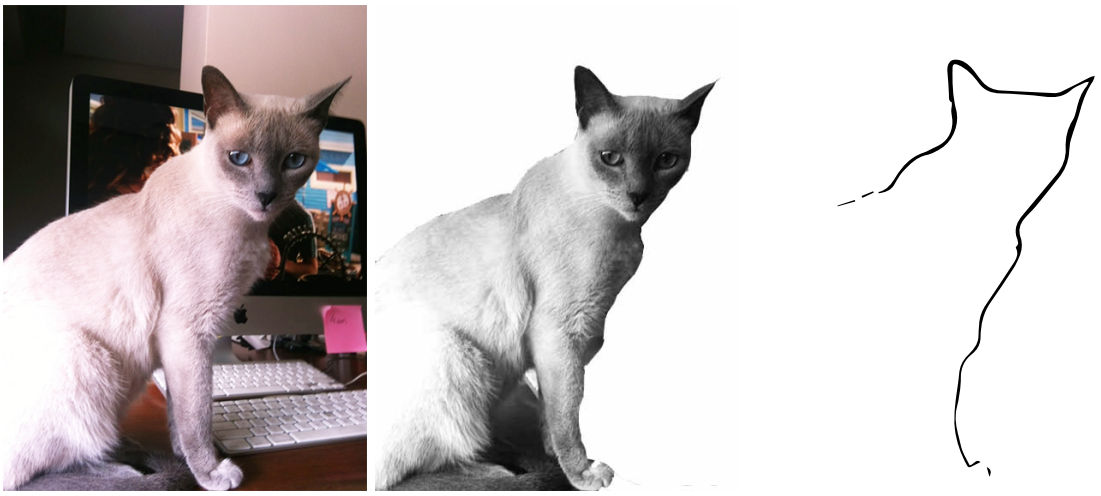
Basically, an order-k ergodic Markov chain is a process with memory of k states. By modeling the source as an ergodic Markov chain, Shannon showed that his theory not only work for phenomena that can be modeled as i.i.d. random variables. The source can behave as a chain of random variables $\{S\}$, each representing an outcome $s \in \mathbb{A}_S$ that are dependent on each other, as long as the sequence produced is longer than the number of symbols needed to the Markovian process achieve its stability.

4.6 THE ENCODER: DATA COMPRESSION

What's in a name?
That which we call a rose,
by any other word would smell as sweet.

– Romeo and Juliet (act.2, sce.2), William Shakespeare

An encoding transforms information in data. The same information can be transformed in an audio file with spoken English, a piece of writing in Portuguese, or even an image. These encodings represent the information uniquely and differ amount themselves in the ammount of data (*bits*) they use.



(a) 360 Kb 15cmx20cm PNG colored image of a cat.

(b) 27 Kb 15cmx20cm JPG grayscale image of a cat.

(c) 4.9 Kb 15cmx20cm SVG duotone image of a cat.

Figure 4.3: Different representations of a cat and their encoding sizes in bits.

This idea may be better explained with an analogy with natural languages. Languages encode ideas into words in different ways [66]. For example, while in English's *"to be"* is universal, Portuguese has two different verbs: *"ser"* and *"estar"*; the first for permanent, unchanging cases; the second for temporary situations such as mood or weather. At the same time, similar or identical meanings appear in unrelated languages [66].

Thus, a message in a natural language can be translated (encoded) to another language and both messages will hardly have the same number of words, characteres, or size in *bits*:

$$S^n = \{S_1, \cdots, S_n\} \xrightarrow{\text{encoding}} \{X_1, \cdots, X_k\} = X^k. \tag{4.20}$$

$$\tag{4.21}$$

Besides, some symbols are more important in a message: "Mst nglsh spkrs wll ndrstnd ths phrs wtht vwls[1]". Here we created *codewords* for words in English that a receiver can understand by the context (and certainly if she has a *codebook*).

A codebook *is a dictionary that relates words in the source alphabet, $\mathbb{A}_S$ to words, codes, in the encoder alphabet $\mathbb{A}_X$.*

Shannon's source coding theorem is about encoding messages efficiently, a form of data compression [57]. Here we present some definitions that will help us understand the theorem later.

1 "Most English speakers will understand this phrase without vowels".

Definition 21. *A **(m, k) block code**, also known as a codebook, is a set of m codewords represented by a sequence of k bits:*

$$\{X^k(1), X^k(2), ..., X^k(m)\},\ X^k(i) \in \mathbb{A}_X^k,\ m \in \mathbb{N}. \tag{4.22}$$

Definition 22. *The rate $\mathbb{R}_{code}$ of an (m,k) code is:*

$$\mathbb{R}_{code} = \frac{\log m}{k}\ \frac{bits}{usage} \tag{4.23}$$

Definition 23. *Let S^n be a block of n random variables, representing consecutive symbols $S_i \in \mathbb{A}_S$ emited by the source. A **binary block encoder** X is a function:*

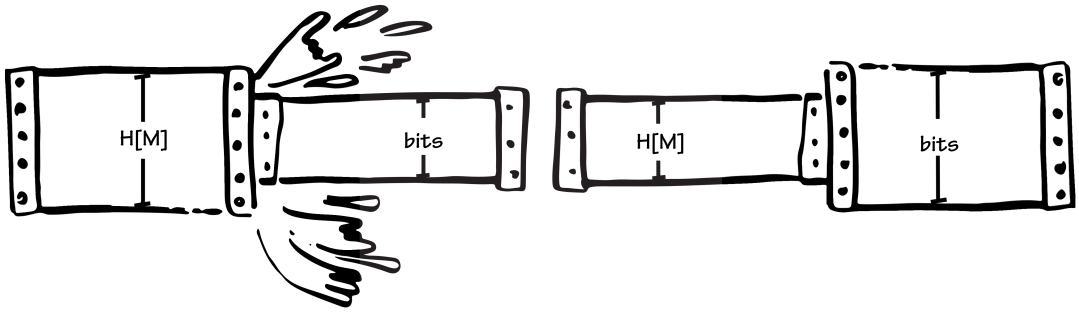
$$X:\ \mathbb{A}_S^n \to \{0,1\}^k \tag{4.24}$$

that "translates" the block of source symbols (the message) into a code X^k of k bits, using a $(|\mathbb{A}_S^n|, k)$ code:

$$X(S^n) = \{x_1, \cdots, x_k\} = \boldsymbol{x} \in \{0,1\}^k \tag{4.25}$$

Definition 24. *The rate $\mathbb{R}_X$ of a binary block encoder is:*

$$\mathbb{R}_X = \frac{\log |\mathbb{A}_S^n|}{k} = \frac{n}{k} \log |\mathbb{A}_S|\ \frac{bits}{usage} \tag{4.26}$$



(a) Loss of information. **(b)** Waste of resources (bits).

Figure 4.4: Entropy of the source vs. coding capacity.

Shannon proved that there is a relation between the entropy of the source and its optimal encoding (this relation will be shown in section 4.6.6). The entropy of the source is a lower bound on the

minimum bits/symbol needed to encode it. The intuition is simple, imagine the entropy of the source as a "tube" (see figure 4.4). The capacity of the tube is the rate of bits/symbol we expect from the source. The encoder is a connection to the tube.

If we use less bits than the entropy to encode it, we are losing information (see figure 4.4a). If we use more bits than the entropy, we are wasting resources, *bits* (see figure 4.4b).

4.6.1 An encoding example

Let us use an example to better illustrate this crucial concept in IT[2]. Imagine you are building a weather station that sends the moment weather condition to a distant control room. Also, there are 8 weather conditions we are interested in. In this case, a message is the transmission of one symbol from $\mathbb{A}_S$.

$$\mathbb{A}_S = \{w_0, w_1, w_2, w_3, w_4, w_5, w_6, w_7\} \tag{4.27}$$

How can we encode these weather conditions?

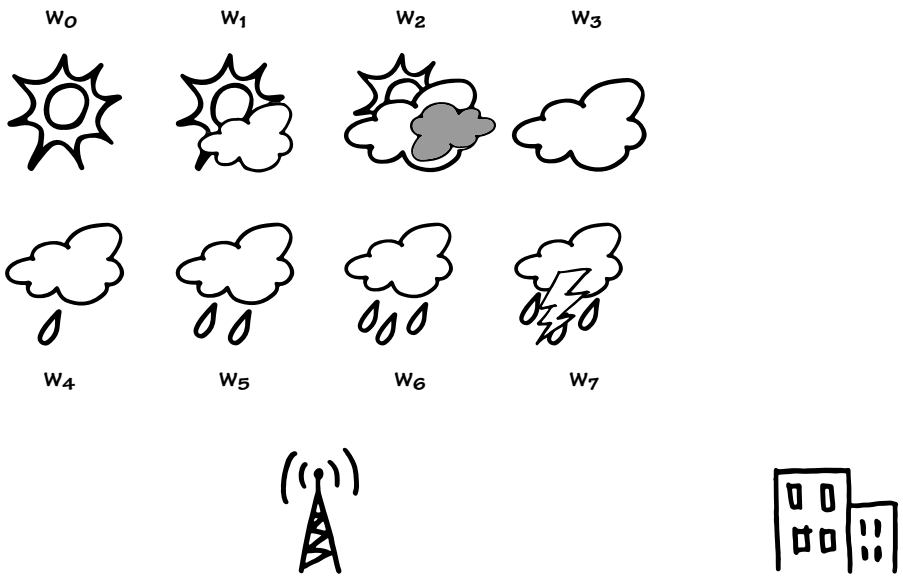


Figure 4.5: A weather station.

4.6.2 Raw bit content

The first idea is to enumerate $\mathbb{A}_S$ in binary, using 3 bits/symbol.

$$\begin{aligned}\mathbb{A}_X = \{&x_0 = 000, x_1 = 001, x_2 = 010, x_3 = 011,\\ &x_4 = 100, x_5 = 101, x_6 = 110, x_7 = 111\}\end{aligned} \tag{4.28}$$

This encoding provide a model of the source that has maximum

2 This example is inspired by Géron [22]

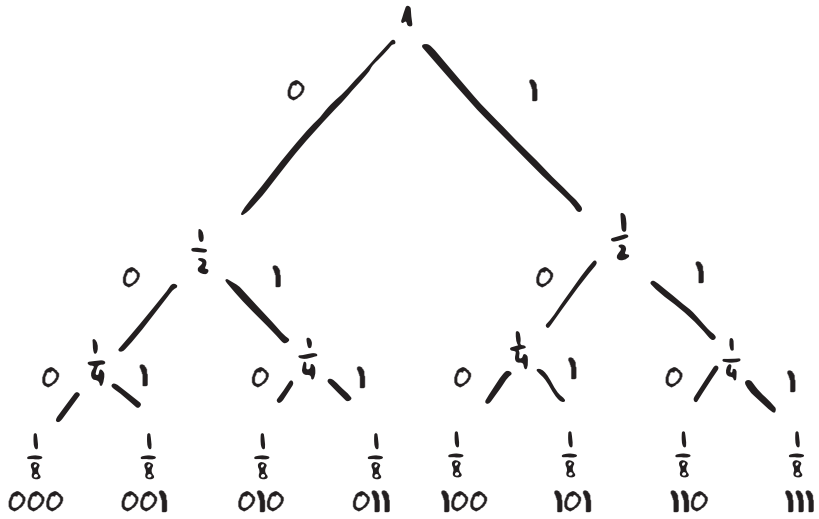


Figure 4.6: Largest encoding = Maximum entropy.

entropy (all outcomes are equiprobable, thus have the same encoding size):

The probability distribution that produces maximum entropy is the uniform distribution *(section 3.11.1)*

$$p(x_i) = \frac{1}{|\mathbb{A}_X|}, \forall i \in [0, 7] \tag{4.29}$$

$$H[X] = -\sum^{|\mathbb{A}_X|} \frac{1}{|\mathbb{A}_X|} \log \frac{1}{|\mathbb{A}_X|} \tag{4.30}$$

$$= \log |\mathbb{A}_X|. \tag{4.31}$$

Is this a good encoding?

4.6.3 Maximum Entropy Principle

If all information we have is how many weather conditions are there, the size of the source alphabet, the best model is the one which conveys this information and has maximum entropy, i. e. it makes no further assumptions. This is the model where we have the worst case scenario for the average number of questions needed to find out which outcome is the right one:

$$P_S = \{p_0 = \tfrac{1}{8}, p_1 = \tfrac{1}{8}, p_2 = \tfrac{1}{8}, p_3 = \tfrac{1}{8}, \\ p_4 = \tfrac{1}{8}, p_5 = \tfrac{1}{8}, p_6 = \tfrac{1}{8}, p_7 = \tfrac{1}{8}\} \tag{4.32}$$

So, in this case that encoding (equation 4.28) is indeed a good option. Notice that the encoding process yields a specific distribution $p(X)$, which determines its entropy $H[X]$ and, therefore, how much information per symbol it carries [57]. The maximum entropy is obtained with this equiprobable distribution, the *uniform distribution* (section 3.11.1).

Let us assume now that another information about the source is given. The weather station is in Atacama, and $P_{M'} = \{p_0 = 75\%, p_1 =$

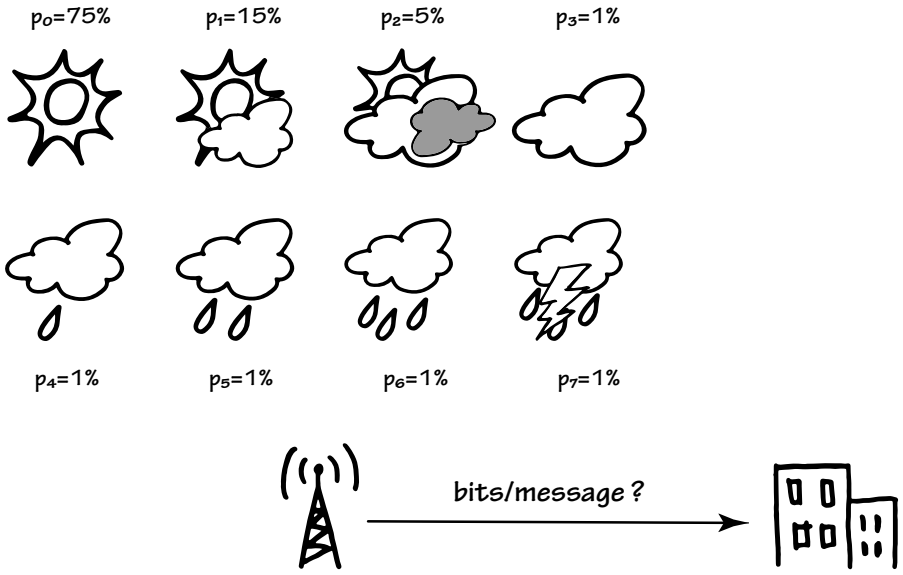


Figure 4.7: A weather station in Atacama.

$10\%, p_2 = 5\%, p_3 = 1\%, p_4 = 1\%, p_5 = 1\%, p_6 = 1\%, p_7 = 1\%\}$. With this new information about the source. Can we do better? Sure.

First, let us calculate the lower bound (maximum efficiency) of the bits/symbol rate of the source encoding:

$$\begin{aligned} H[S'] &= 0.75 \log \frac{1}{0.75} + 0.15 \log \frac{1}{0.15} + 0.05 \log \frac{1}{0.05} + 5\left(0.01 \log \frac{1}{0.01}\right) \\ &\approx 1 \frac{\text{bit}}{\text{message}} \end{aligned} \tag{4.33}$$

We know that theoretically we cannot have an encoding with less than 1 bit/symbol in average. But we can improve from 3 bits/symbol (see figure 4.8):

Any distribution that is not uniform will lead to an average tree height that is smaller that the uniform distribution. The uniform distribution is the worst case.

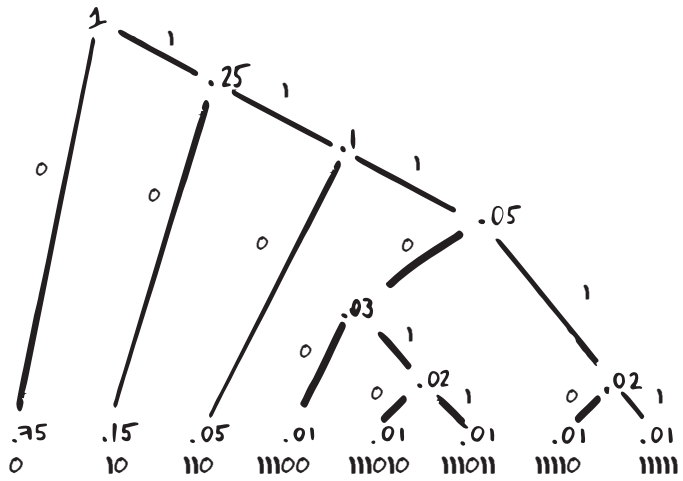


Figure 4.8: The probability distribution of the source determines an encoding.

$$\begin{aligned}\mathbb{A}_{X'} = \{&x'_0 = 0, x'_1 = 10, x'_2 = 110, x'_3 = 11100,\\ &x'_4 = 111010, x'_5 = 111011, x'_6 = 11110, x'_7 = 11111\}\end{aligned} \tag{4.34}$$

The average encoding size per message symbol in X' is:

$$\begin{aligned}&0.75 \cdot 1 + 0.15 \cdot 2 + 0.05 \cdot 3 + 0.03 \cdot 5 + 0.02 \cdot 6\\ &\approx 1.5 \frac{\text{bits}}{\text{symbol}}\end{aligned} \tag{4.35}$$

4.6.4 Cross Entropy

This average encoding size per message has a special name: the Cross Entropy. It is evident the similarity of the definition of Cross-Entropy and Entropy. If our model q of the real distribution p is absolute right, the Cross-Entropy is equal the Entropy $H_{p,q} = H_p$. If not (as it is in most cases), $H_{p,q} > H_p$.

In our Atacama weather station example, the cross-entropy between the real distribution $p = p(s)$ and the encoding distribution $q = p(x)$ was 1.5 bits/symbol. So, we can say the efficiency of the encoding $X(s)$ is $\frac{\text{information}}{\text{data}} = \frac{H[S]}{H_{p,q}[S]} = \frac{1}{1.5} \approx 67\%$. We calculated $H_{p,q}$ knowing the sizes of each possible s_i.

Let us use another example, imagine that we transport the weather station from Atacama to London, where the probability distribution of the weather is $P_{S''} = \{p_0 = 5\%, p_1 = 5\%, p_2 = 10\%, p_3 = 15\%, p_4 = 15\%, p_5 = 20\%, p_6 = 20\%, p_7 = 10\%\}$, and keep using the same encoding. It is obvious that the encoding will me much less efficient. The

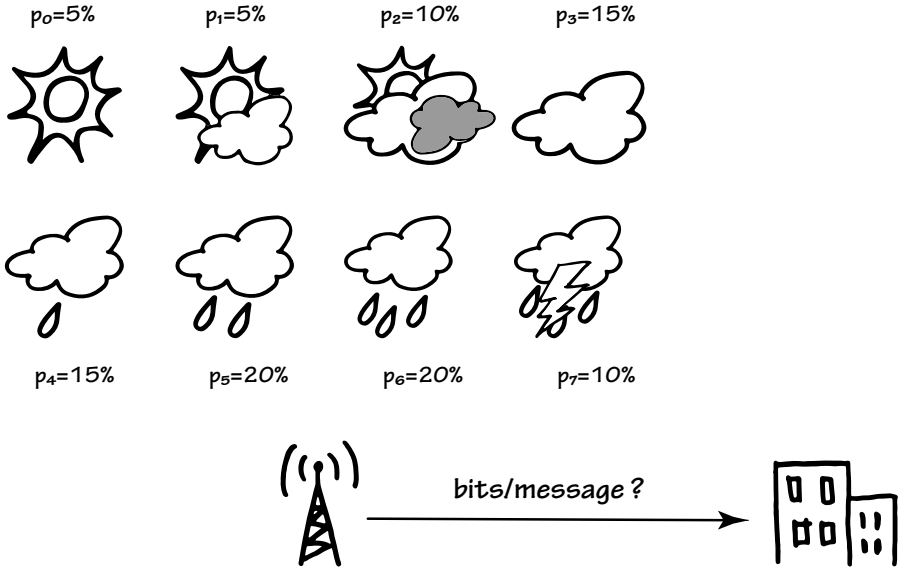


Figure 4.9: The Atacama's weather station in London.

average size of a message symbol in this situation is:

$$\begin{aligned} H_{p,q}[X'],\, p = P(S''),\, q = P(X') & \qquad (4.36)\\ = 0.05 \cdot 1 + 0.05 \cdot 2 + 0.1 \cdot 3 + 0.45 \cdot 5 + 0.35 \cdot 6 & \\ \approx 4.8 \frac{\text{bits}}{\text{symbol}} & \qquad (4.37) \end{aligned}$$

Definition 25. ***Cross entropy*** *is the average number of bits needed to encode data coming from a source X with distribution $p(x)$ when using model $q(x)$.*

$$H_{p,q}[X] = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \qquad (4.38)$$

4.6.5 KL Divergence (or Relative Entropy)

The amount by which the Cross-Entropy and the Entropy diverge is the KL Divergence:

Definition 26. *The* ***relative entropy or Kullback–Leibler divergence*** *between two probability distributions $p(x)$ and $q(x)$ that are defined over the same alphabet $\mathbb{A}_X$ is:*

$$D_{KL}(p||q) = \sum_x p(x)\ \log \frac{p(x)}{q(x)} \qquad (4.39)$$

$$D_{KL}(p||q) = H_{p,q}[X] - H_p[X] \qquad (4.40)$$

In our example:

$$D_{KL}(p_{\text{Atacama}}||q_{\text{London}}) = H_{p,q}[X] - H_p[X] \approx 3.3 \frac{\text{bits}}{\text{symbol}} \qquad (4.41)$$

4.6.6 Shannon's source encoding theorem

Now that we understand how the source encoding works, let us take a moment to appreciate the genius of Shannon. Here, we show how he demonstrated the size of the optimal encoding without ever explaining which encoding is that in the first place.

Theorem 6. *The optimal binary enconding $X^k = (X_1, \cdots, X_k)$, $X_i \in \{0, 1\}$, of a n-symbols message $S^n = (S_1, \cdots, S_n)$, where $S_i \in \mathbb{A}_S$ are i.i.d.*[3] *$\sim p(s)$ has size $k \approx nH[S]$ for large n.*

3 An ~~obsessive~~ observant reader may have noticed that we are here considering the source as an i.i.d. stochastic process, instead of a stationay ergodic process. This is the same proof stated by Shannon in [48] and in several IT books (e. g. [14, 34]). A proof for ergodic finite alphabet sources can be found in [39].

Proof. A one-to-one mapping $S^n \mapsto X^k$ is invertible. If we enumerate all elements of S^n in binary, we will need k bits. Thus, with absolute certainty:

$$k \leqslant \log\lceil |S^n| \rceil = \log\lceil 2^{n \log |\mathbb{A}_S|} \rceil = n \log |\mathbb{A}_S| + 1 \text{ bits} \tag{4.42}$$

Can we do better? We know from statistics that most possible outcomes are unlikely. In other words, there is a small set of very likely outcomes that is most probable. So let us use this property of Nature.

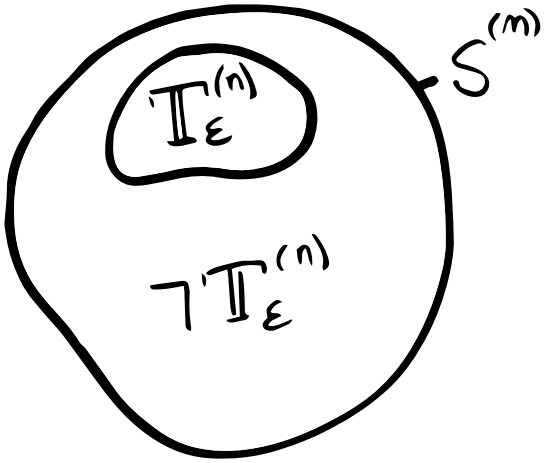


Figure 4.10: The typical set of sequences S^n.

We will divide all sequences S^n into two sets: the typical set ($\mathbb{T}_\epsilon^{(n)}$) and its complement, the atypical set ($\neg\,\mathbb{T}_\epsilon^{(n)}$), which can be seen in figure 4.10.

Definition 27. *The **typical set** $\mathbb{T}_\epsilon^{(n)}$ with respect to $p(s)$ is the subset of sequences $S^n = (S_1, \cdots, S_n), S_i \in \mathbb{A}_S$, where:*

$$P(\mathbb{T}_\epsilon^{(n)}) = \sum_{S^n \in \mathbb{T}_\epsilon^{(n)}} P(S^n) > 1 - \epsilon, \text{ for sufficiently large } n. \tag{4.43}$$

In other words, for a sequence of n i.i.d random variables $S \equiv (S_1, \cdots, S_n)$, each drawn from $p(s)$, the outcome $\boldsymbol{m} = (s_1, \cdots, s_n)$ is almost certain to belong to the typical set $\mathbb{T}_\epsilon^{(n)}$, if n is large.

Let us put aside for a moment that we do not know the size of the typical set, $|\mathbb{T}_\epsilon^{(n)}|$.

We know that:

$$|\mathbb{T}_\epsilon^{(n)}| \ll |\neg\mathbb{T}_\epsilon^{(n)}| < |S^n|, \tag{4.44}$$

$$P(\mathbb{T}_\epsilon^{(n)}) \gg P(\neg\mathbb{T}_\epsilon^{(n)}), \tag{4.45}$$

$$\mathbb{E}(k) = \lceil P(\mathbb{T}_\epsilon^{(n)}) \log |\mathbb{T}_\epsilon^{(n)}| + P(\neg\mathbb{T}_\epsilon^{(n)}) \log |\neg\mathbb{T}_\epsilon^{(n)}| \rceil. \tag{4.46}$$

Therefore, from equation 4.42 we can predict that:

$$\mathbb{E}(k) \ll n \log |\mathbb{A}_S| + 1 \text{ bits} \tag{4.47}$$

Now, we need to find $|\mathbb{T}_\epsilon^{(n)}|$. For this we will use the Asymptotic Equipartition Property (AEP), formalized in the following theorem [14]:

Theorem 7 (AEP). *If $S_1, \cdots, S_n$ are i.i.d. $\sim p(s)$, then:*

$$-\frac{1}{n} \log p(S_1, \cdots, S_n) \to H[S] \text{ in probability.} \tag{4.48}$$

Proof. From the theorem definition, S_i are independent. Then:

$$-\frac{1}{n} \sum_{i=1}^{n} (p(\underbrace{S_1, \cdots, S_n}_{S^n}) = -\frac{1}{n} \log(\prod_{i=1}^{n} p(S_i)) \tag{4.49}$$

$$\overset{\text{eq.3.14}}{=} \frac{1}{n} \sum_{i=1}^{n} -\log p(S_i) \tag{4.50}$$

From the weak law of large numbers:

$$n \to \infty, \frac{1}{n} \sum_{i=1}^{n} \xi_i \to \mathbb{E}(\xi) \tag{4.51}$$

Therefore, using the fact that a statistic of a random variable is a random variable, let $\xi = -\log p(S_i)$ [14] and using equation 4.15 and equation 4.51:

$$n \to \infty,$$

$$\frac{1}{n} \sum_{i=1}^{n} (-\log p(S_i)) \to \underbrace{\mathbb{E}_p(-\log p(S))}_{H[S]} \tag{4.52}$$

$$\therefore -\frac{1}{n} \log p(S^n) \to H[S] \quad \square \tag{4.53}$$

$\square$

Now that we proved the AEP theorem (theorem 7), let us use it to define $|\mathbb{T}_\epsilon^{(n)}|$:

$$-\frac{1}{n} \log p(S^n) \to H[S] \text{ in probability} \tag{4.54}$$

$$p(S^n) \to 2^{-n(H[S])} \therefore \tag{4.55}$$

$$2^{-n(H[S]+\epsilon)} \leqslant p(S^n) \leqslant 2^{-n(H[S]-\epsilon)} \text{ in probability} \tag{4.56}$$

We also now that:

$$1 = \sum_{S^n} p(S^n) \tag{4.57}$$

$$1 \geqslant \sum_{S^n \in \mathbb{T}_\epsilon^{(n)}} p(S^n) \tag{4.58}$$

$$1 \geqslant |\mathbb{T}_\epsilon^{(n)}| \, p(S^n) \tag{4.59}$$

From equation 4.56:

$$1 \geqslant |\mathbb{T}_\epsilon^{(n)}|\, 2^{-n(H[S]+\epsilon)} \tag{4.60}$$

$$\therefore\ |\mathbb{T}_\epsilon^{(n)}| \leqslant 2^{n(H[S]+\epsilon)} \tag{4.61}$$

This upper bound to $|\mathbb{T}_\epsilon^{(n)}|$ is all we need to prove source coding theorem (theorem 6).

$$\begin{aligned}
\mathbb{E}(k) &= \lceil P(\mathbb{T}_\epsilon^{(n)}) \log |\mathbb{T}_\epsilon^{(n)}| \\
&\quad + \cancelto{\epsilon}{P(\mathbb{T}_\epsilon^{(n)})} \cancelto{|S^n| = n\log|\mathbb{A}_S|}{\log|\mathbb{T}_\epsilon^{(n)}|} \rceil && (4.62)\\
&\simeq \lceil (1-\epsilon)\log 2^{n(H[S]+\epsilon)} + \cancelto{\epsilon' n}{\epsilon n \log|\mathbb{A}_S|} \rceil && (4.63)\\
&\simeq \lceil (1-\epsilon)[n(H[S]+\epsilon)] + \epsilon' n \rceil && (4.64)\\
&\simeq \lceil n(H[S]+\epsilon-\epsilon n H[S]-\epsilon^2)] + n(\epsilon') \rceil && (4.65)\\
&\simeq \lceil n(H[S]+\cancelto{\epsilon''}{\epsilon-\epsilon H[S]-\epsilon^2})] + n(\epsilon') \rceil && (4.66)\\
&\simeq \lceil n(H[S]+\epsilon''+\epsilon') \rceil = \lceil n(H[S]+\varepsilon) \rceil && (4.67)
\end{aligned}$$

$$\therefore$$

$$\mathbb{E}(k) \simeq nH[S] \qquad \square$$

We proved that the average information per symbol of the coding generated by the optimum encoder has the same average information per symbol as the source, $H[S]\frac{\text{bits}}{\text{symbol}}$. Due to this property, it is quite common to talk about $H[X]$ as the entropy of the source.

4.6.7 Typical Set

In the proof of the source coding theorem, we defined the typical set and discovered some of its properties, but we left one behind. We only needed the upperbound of the $|\mathbb{T}_\epsilon^{(n)}|$, let us now derive its lowerbound. From equation 4.56 and the typical set definition (equation 4.43):

$$\sum_{S^n \in \mathbb{T}_\epsilon^{(n)}} 2^{-n(H[S]-\epsilon)} \geqslant 1-\epsilon \tag{4.68}$$

$$|\mathbb{T}_\epsilon^{(n)}| 2^{-n(H[S]-\epsilon)} \geqslant 1-\epsilon \tag{4.69}$$

$$|\mathbb{T}_\epsilon^{(n)}| \geqslant (1-\epsilon)2^{n(H[S]-\epsilon)} \tag{4.70}$$

Therefore, from equation 4.70 and equation 4.61 we can derive:

$$(1-\epsilon)2^{n(H[S]-\epsilon)} \leqslant |\mathbb{T}_\epsilon^{(n)}| \leqslant 2^{n(H[S]+\epsilon)} \tag{4.71}$$

$$|\mathbb{T}_\epsilon^{(n)}| \to 2^{nH[S]} \tag{4.72}$$

With that, we can list some useful properties of $\mathbb{T}_\epsilon^{(n)}$:

1. almost all probability is concentrated in the typical set, by definition (equation 4.43);

2. elements in the typical set are nearly equiprobable, equation 4.56;
3. the number of elements in the typical set is nearly $2^{H[S]}$, equation 4.72.

Going back to the AEP theorem (theorem 7):

$$\begin{aligned} \frac{1}{n}\log\left(\frac{1}{p(S^n)}\right) &\to H[S] \\ H[S]-\epsilon \leqslant \frac{1}{n}\log\left(\frac{1}{p(S^n)}\right) &\leqslant H[S]+\epsilon \end{aligned} \tag{4.73}$$

We can think of the middle term as the entropy of a sample os size n, thus a typical sample give us an amount of information close to the average information from the source, $H[S]$.

This insight reminds us of the sample complexity which will be discussed in chapter 5

4.7 THE CHANNEL: DATA TRANSMISSION

The channel is simply the medium used to transmit the signal $\boldsymbol{x}$ from the encoder to the decoder. It may be anything from a band of radio frequencies, an electrical wire, a beam of light, or a postal service. As we did before, we can also think the channel as a "tube" which carries information (see figure 4.4).

This definition of a discrete channel covers the deterministic case where $y = f(x)$. In most of the cases the usage of channel is determined by a period in which it is being used. Thus, some prefer to define the capacity in bits/second.

Definition 28. *Mathematically, a **discrete channel** is the conditional probability*

$$p(y|x), y \in \mathbb{A}_Y, x \in \mathbb{A}_X. \tag{4.74}$$

4.7.1 Noiseless Channel Capacity

Definition 29. *The **operational capacity** of a channel is the maximum rate of bits per usage that the medium is physically capable of transmitting. It is in fact just a number of bits per usage. We can think of it as the maximum entropy it is capable of transmiting in the absence of noise:*

$$\mathbb{C}_{operational} = R = \max_{p(x)} \log|\mathbb{A}_X| \ bits/usage. \tag{4.75}$$

4.7.2 The noisy channel

All practical communications, however, are noisy [57]. Noise reduces the rate at wich information can be communicated reliably. Shannon

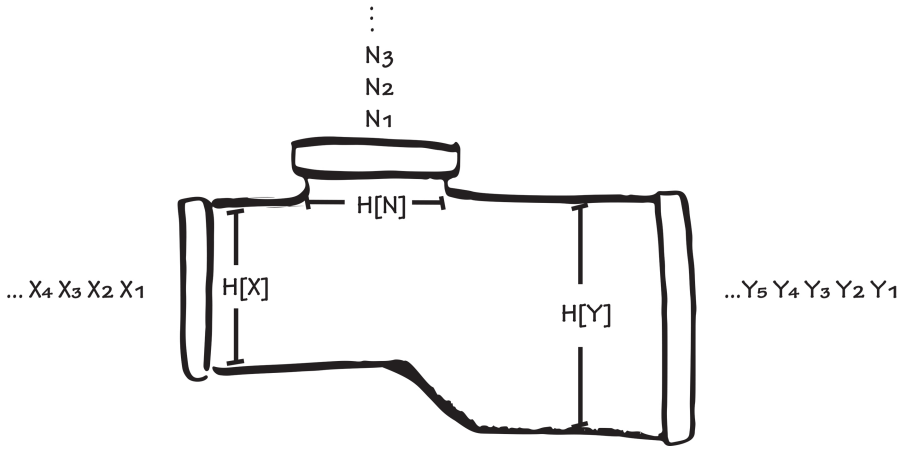


Figure 4.11: The noisy channel.

proved that information can be communicated, with arbitrarily small error, at a rate which is limited only by the channel capacity.

To understand how noise affects the channel capacity, we need to understand the concepts of **conditional entropy**, **joint entropy** and **mutual information**.

4.7.3 Conditional Entropy

The residual uncertainty we have about a random variable Y, given that we already know the outcome of another random variable X is the **conditional entropy**:

Definition 30. *The **conditional entropy** or **equivocation** $H[X|Y]$ of X given Y is:*

$$H[X|Y] \triangleq \sum_{y \in \mathbb{A}_Y} p(y) \left[\sum_{x \in \mathbb{A}_X} p(x|y) \log \frac{1}{p(x|y)} \right] \tag{4.76}$$

$$= - \sum_{xy \in \mathbb{A}_X \mathbb{A}_Y} p(x, y) \log p(x|y) \tag{4.77}$$

4.7.4 Joint Entropy

We have defined the entropy of a single random variable in equation 4.15. Now, we extend the definition to a pair of random variables. As the pair can be seen as a single vector-valued random variable, there is nothing new in this definition [14, p.15].

Definition 31. *The **joint entropy** $H[X, Y]$ of a pair of discrete random variables (X, Y) with joint distribution $p(x, y)$ is defined as:*

$$H[X, Y] \triangleq -\mathbb{E} \log p(X, Y) \tag{4.78}$$

$$= -\sum_{x \in \mathbb{A}_x} \sum_{y \in \mathbb{A}_y} p(x, y) \log p(x, y). \tag{4.79}$$

4.7.5 Mutual Information

Definition 32. *The **mutual information** $I[X; Y]$ between two variables, such as a channel input X and output Y, is the amount of information obtained about one random variable through observing the other random variable.*

$$I[X; Y] = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \text{ bits} \tag{4.80}$$

$$= H[X] - H[X|Y] \tag{4.81}$$

$$= H[Y] - H[Y|X] \tag{4.82}$$

$$= H[X] + H[Y] - H[X, Y] \tag{4.83}$$

$$= H[X, Y] - [H[X|Y] + H[Y|X]] \text{ bits} \tag{4.84}$$

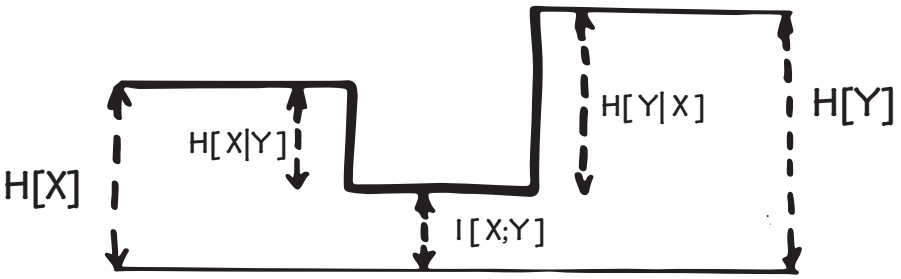


Figure 4.12: Relationship between information measures in a channel.

For a visual understanding of this measures, see figure 4.12. The mutual information can also be seen as a measure of the mutual dependence between the two variables, as the mutual information is the same as the Kullback–Leibler divergence between the joint distribution and the product of the variables marginal distributions:

$$I[X; Y] = D_{KL}(p(x, y) || p(x)p(y)). \tag{4.85}$$

4.7.6 Noisy channel capacity

Given that in a noise channel $Y = X + N$, where N is the noise in the channel, from the mutual information definition:

$$\begin{aligned} I[X;Y] &= H[Y] - H[Y|X] && (4.86)\\ &= H[Y] - H[(X+N)|X]. && (4.87)\end{aligned}$$

If X is known, the uncertainty from X is none:

$$I[X;Y] = H[Y] - H[N|X] \quad (4.88)$$

By definition, N and X are independent, therefore:

$$\begin{aligned} I[X;Y] &= H[Y] - H[N] && (4.89)\\ \therefore H[Y|X] &= H[N] && (4.90)\end{aligned}$$

Definition 33. *The **information capacity** or* effective capacity *of a noisy channel is defined as:*

$$\begin{aligned} C &= \max_{p(x)} I[X,Y] && (4.91)\\ &= \max_{p(x)} (H[Y] - H[Y|X]) \textit{ bits/usage.} && (4.92)\\ &= \max_{p(x)} (H[X] - H[X|Y]) \textit{ bits/usage.} && (4.93)\end{aligned}$$

The information capacity can be derived theorem from Shannon's noisy channel theorem (4.8.1). Notice that when there is no noise, the definition corresponds to the *operational capacity* definition.

4.8 THE DECODER

The decoder capability of reconstructing the message from the encoded data despite the channel noise is a direct consequence of Shannon's second theorem:

4.8.1 Shannon's noisy channel theorem

In his second and, pehaps, most important theorem, Shannon proved that, provided $H[X] \leqslant C$, the average error (ϵ), when averaged over all possible encoders, approaches to zero ($\epsilon \to 0$) as the length of the input $\boldsymbol{x}$ increases. Therefore, there must exist at least one encoder that produces an error as small as ϵ.

Once again, Shannon proved the counterintuitive argument that there is an encoder that produces an arbitrarily small error without showing how to find this encoder.

Instead of proving the theorem (for which we refer to MacKay [34] and Cover and Thomas [14]), let us give an intuitive preview of the proof.

Consider n uses of the channel as our block usage. There are $|\mathbb{A}_X|^n$ possible inputs $\boldsymbol{x}$ and $|\mathbb{A}_Y|^n$ possible outputs $\boldsymbol{y}$ in the block usage. We want to prove that for any $\boldsymbol{y}$, it is possible to derive an unique message that generated it.

If n is large, any particular $\boldsymbol{x} \in X^n$ is very likely to produce an output in a small subspace of the output alphabet, the typical output set, given $\boldsymbol{x}$. So, it is possible to find a non confusable subset of the input sequences that produces disjoint output sequences (figure 4.13).

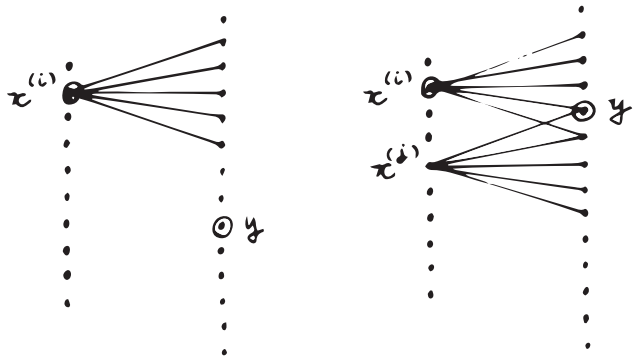


Figure 4.13: Disjoint output sequences.

Take $\boldsymbol{x} \sim p(X^n)$. Recall the source coding theorem (theorem 6), the total number of typical output sequences $\boldsymbol{y}$ is $2^{nH[Y]}$, all sequences being almost equiprobable. For any sequence $\boldsymbol{x}$, there are about $2^{nH[Y|X]}$ probable sequences. Now we restric ouselves to the subset of the typi-

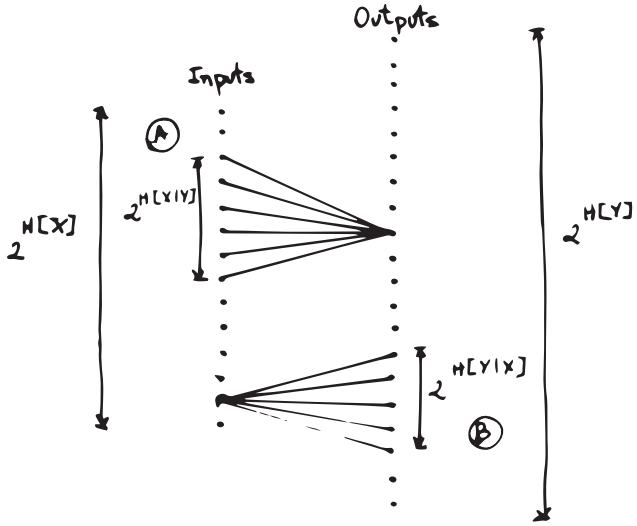


Figure 4.14: The need to restrict to the subset of typical inputs and its implications.

cal inputs such that the correponding typical output sets are disjoint. We can expect the number of non-confusable inputs to be:

$$\leqslant \frac{2^{nH[Y]}}{2^{nH[Y|X]}} = 2^{n(H[Y]-H[Y|X])} = 2^{nI[X;Y]} \tag{4.94}$$

The maximum value of this bound is achieved by the process X that maximises $I[X;Y]$. Therefore, $n \max_{p(x)} I[X;Y]$ is the maximum amount of bits that can be transmited in n usages of the channel, which proves the first law of information (see section 4.3):

$$C_{\text{noisy channel}} = \max_{p(x)} I[X,Y]. \tag{4.95}$$

We can rewrite equation 4.95 as:

$$C_{\text{noisy channel}} = \max_{p(x)} (H[X] - H[N]), \tag{4.96}$$

wich states that noise reduces channel capacity. So, this is also a proof for the second law of information (section 4.3).

5 MACHINE LEARNING THEORY

Mathematics operates inside the thin layer
between the trivial and the intractable.

– Andrey Kolmogorov

In wich we present the theoretical framework of Machine Learning, the PAC model, theoretical guarantees for generalisation, and expose criticism due to its lack of explanation on Deep Learning phenomena[1].

5.1 MOTIVATION

As already discussed, learning is the process of inferring general rules to perform a specific task by observing limited examples. The learning algorithm must not only perform well in the sample already seen but, more importantly, in previously unseen examples.

How can we prove that an algorithm learned? We may know its performance in the given sample, but does it translate to any sample? Can we guarantee bounds to the error in an unknown distribution of examples even if we have just a limited sample of it? Can we bound the number of samples needed (sample complexity) to obtain a sure accuracy on unseen examples? How does the sample complexity grow? These are the kind of questions that motivated the development of Machine Learning Theory (MLT). This research field started in Russia by the name of Statistical Learning Theory (SLT), during the late 1960_s, with the work of Vapnik and Chervonenkis (see figure 5.1). In 1984, Leslie Valiant proposed the Probably Approximately Correct (PAC) framework to bring ideas from the Computational Complexity Theory to learning problems, giving birth to the field of Computational Learning Theory (CoLT). Up to the 1990_s, STL and CoLT were very active fields.

1 This chapter is influenced by the online lecture Shawe-Taylor and Rivasplata [50], the online lecture series Mello [40] and the book Mello and Ponti [41].

[version: June 19, 2020 at 18:52]

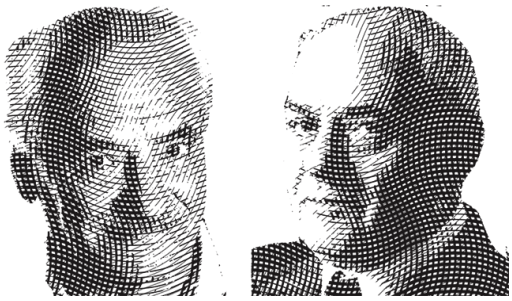


Figure 5.1: Chervonenkis (Left) and Vapnik (Right).

Some will say CoLT is a subfield of SLT, others just the contrary. SLT and CoLT use similar mathematical analysis, with only some different jargon. It can be argued that they differ on their objectives: CoLT focuses on classifying learning problems; SLT focuses on analysing and improving the accuracy of learning algorithms. The difference, however, seems to be less of substance and more social.

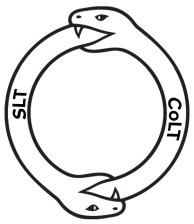


Figure 5.2: SLT and CoLT.

This document tries to be out of this feud and consider both as fundamental Machine Learning Theory. We will also limit our overview of MLT to the context of supervised binary classification problems. This limitation is not a deficiency of the theory, but a mere choice of scope for this document.

5.2 THE LEARNING PROBLEM

The goal of learning is coming up with an understanding of the world from experience, a theory, a tested hypothesis.

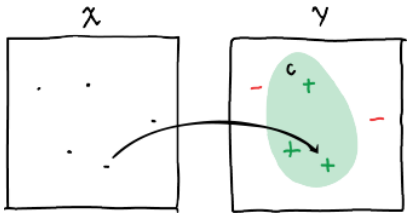


Figure 5.3: A *concept* c is an idealised mapping from the input space $\mathcal{X}$ to output space $\mathcal{Y}$. Each point on the left is an instance x_i. Note that given that $\mathcal{Y}$ is binary, $c \subset \mathcal{X}$.

The learning task is the *concept* c we want to learn. The concept is the idealised function that maps an instance of the problem x_i from the input space $\mathcal{X}$ (also known as problem space) to a solution y_i of the output space $\mathcal{Y}$ (also known as label space). In the PAC framework,

the convention is that labels are binary, $\mathcal{Y} = \{-, +\}$, therefore, $c \subset \mathcal{X}$. A concept class $\mathbb{C}$ is a set of concepts c, thus, a set of sets of $\mathcal{X}$.

We imagine there is a certain distribution $D = P(X, Y)$ in nature, from which $P(X)$, the distribution of examples, and $P(Y|X)$, the learning task, derive. Even knowing nothing about D, we want to discover $P(Y|X)$, given a sample of $(x, y) \sim P(X, Y)$.

5.2.1 The learning problem setting

Supervised learning has three main components (see figure 5.4):

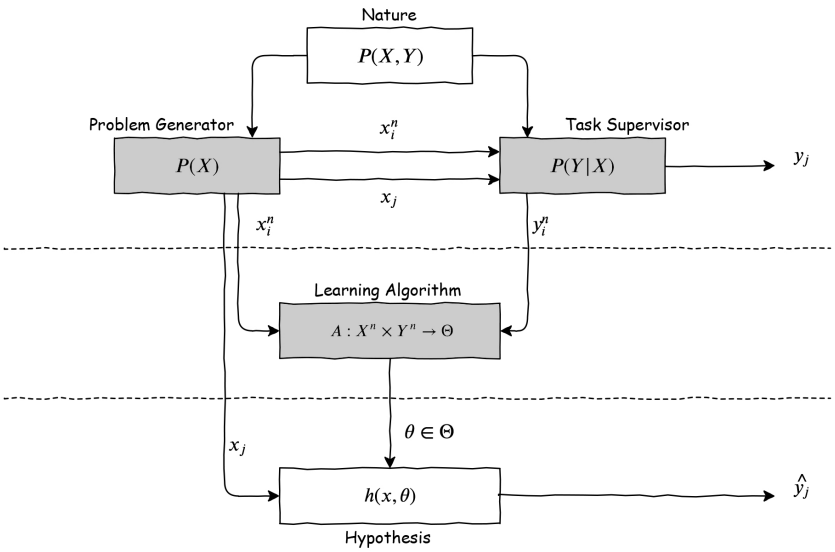


Figure 5.4: Problem setting.

1. A **generator** of random vectors $x \sim P(X), x \in \mathcal{X}$, which represent instances of the problem[2];
2. A **task supervisor** that knows the concept and returns an output vector y_i for every input vector x_i:

$$y_i = c(x_i) = P(Y = y_i \mid X = x_i). \tag{5.1}$$

3. A **learning algorithm** $\mathcal{A}$ which is the functional that given a sample of ninputs and n outputs of a task, $S_n = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, selects an hypothesis h from the hypothesis space $\mathcal{H}$:

$$\mathcal{A} : \underbrace{(\mathcal{X} \times \mathcal{Y})^n}_{S_n} \to \mathcal{H}. \tag{5.2}$$

2 $P(X = x_i) = P_X(x_i) = \sum_j P_{XY}(x_i, y_j) \therefore P_{\mathcal{X}}$is just a consequence of $P(X, Y) \therefore x \sim P_{\mathcal{X}} \equiv x \sim P_{XY}$.

The problem of learning is choosing from the *hypothesis space*[3], the one *hypothesis* that best approximates the *concept*. The selection is based on a training set of n i.i.d. observations drawn according to the unknown distribution $D = P(X, Y)$.

5.2.2 Assumptions

In this chapter, we adopt the MLT community notation for the space of possible inputs as $\mathcal{X}$ and ouputs $\mathcal{Y}$. In information theory we used $\mathbb{A}_X$ and $\mathbb{A}_Y$.

The common assumptions are as follows [41, 62]:

i. **No assumption on $D = P(X, Y)$:** it can be any arbitrary joint probability distribution on $\mathcal{X} \times \mathcal{Y}$.

ii. $D = P(X, Y)$ **is unknown at the training stage:** if not, learning would be trivial.

iii. $D = P(X, Y)$ **is fixed:** There is no "time" parameter, meaning that the ordering of examples in the sample is irrelevant .

iv. **Independent sampling:** examples must be sampled in an identically independent manner (i.i.d.).

v. **Labels may assume non-deterministic values:** due to noise or label overlap.

5.2.3 Hypothesis spaces

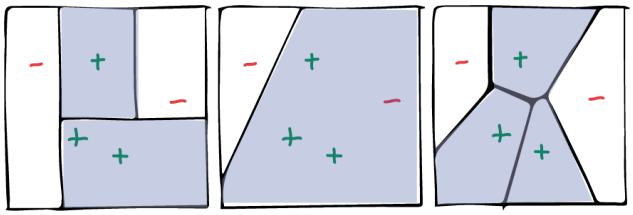


Figure 5.5: Different hypothesis spaces for the same sample.

We can also say that the hypothesis space is the language defined by the learner.

The problem setting rely on the idea of a *hypothesis space* (also known as a *hypothesis class*). A hypothesis space is the set of all hypothesis generatable by a functional learning algorithm $\mathcal{A}$. In the same hypothesis space $\mathcal{H}$, hypotheses are differentiated by their parameter vector θ. Choosing a hypothesis h_i is choosing its parameter θ_i.

$$h : \mathcal{X} \times \Theta \to \mathcal{Y}, \tag{5.3}$$

$$h(x) = p(y \mid x \wedge \theta), \ \theta \in \Theta. \tag{5.4}$$

Different learners will constraint the input space $\mathcal{X}$ differently (see figure 5.5). Some algorithms are more complex than others, meaning they can express more different functions[4].

3 Hypothesis spaces will be explained in section 5.2.3.

4 We also use the term *capacity* to describe this characteristic of learning algorithms to generate more complex hypotheses.

We usually call $\mathcal{H}_{all}$ the hypothesis space of all possible functions. However, generalisation only happens if a learner chooses a subset of $\mathcal{H}_{all}$ where to search for the hypothesis. The need for this constraint in generalisation, a bias, was prooved by Mitchell [42]: "*biases are [...] critical to the ability to classify instances that are not identical to the training instances*". An intuitive argument for this is very simple, if any function was allowed, the learner would be able to choose the function that "memorises" the sample, and that would certainly not generalise to other cases.

5.2.4 Learning as error minimisation

Choosing from the *hypothesis space*, the one *hypothesis* h that **best** approximates the *concept*, which we will call h_{Bayes}, can be seen as an optimisation problem where we want to minimise the error of the approximation:

ABSOLUTE ERROR Let loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a measure of the error between the perfect output y of the supervisor, and the obtained output $\hat{y}$ of the hypothesis. **The risk is the expected loss.** Find θ_* which minimises the risk.

$$R_D(\theta) = \mathbb{E}(\ell(x, y, h(x, \theta)), (x, y) \sim D, \theta \in \Theta \tag{5.5}$$

$$\theta_* = \arg\min_{\theta \in \Theta} R(\theta) \tag{5.6}$$

$$h(x, \theta_*) = h_{\text{Bayes}} = \arg\min_{h \in \mathcal{H}} R(h) \tag{5.7}$$

The risk R_D is also called the absolute (or out-of-sample or theoretical) error of the hypothesis[5]. Nevertheless, there is one crucial caveat: the choice of the loss metric is arbitrary, which curbs any objective, metric independent, interpretation of the results.

EMPIRICAL ERROR The underlying difficulty of risk minimisation is that we are trying to minimise a quantity we cannot evaluate: if $P(X, Y)$ is unknown, we cannot directly compute the risk $R(h)$ (absolute error). However, we can compute the risk of the hypothesis on the training sample:

$$\hat{R}_S(h) = \frac{1}{n} \sum_{i=1}^{n} (\ell(x_i, y_i, h(x_i)), (x, y) \sim S \tag{5.8}$$

With this empirical risk $\hat{R}_S$ that we can evaluate, we find the hypothesis that minimises it. That is, given a sample $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$,

5 R, $R(\theta)$ and $R(h)$ are used interchangeably in this document.

a hypothesis space $\mathcal{H}$, and a loss function ℓ, we define $h_{\mathcal{H}}$ as the function:

$$h_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} \hat{R}_S(h) \tag{5.9}$$

According to the law of large numbers (subsection 5.5.3), if the sample is large enough, by induction, a hypothesis generated optimising $\hat{R}_S$ is close to R. However, it is important to notice we still have to discuss at which rate does $\hat{R}_S$ converge to R w.r.t. the sample size.

5.3 BIAS–VARIANCE TRADE-OFF

When we define a subset of $\mathcal{H} \subset \mathcal{H}_{all}$ where to look for our hypothesis, we are imposing a constraint to the choice, a *bias*. Besides, the subset $\mathcal{H}$ can be larger or smaller. For example, the hypothesis space of Neural Networks is much larger than the one of Perceptrons, and also covers it $\mathcal{H}_{NN} \supset \mathcal{H}_{Perceptron}$. Accordingly, we can distinguish two

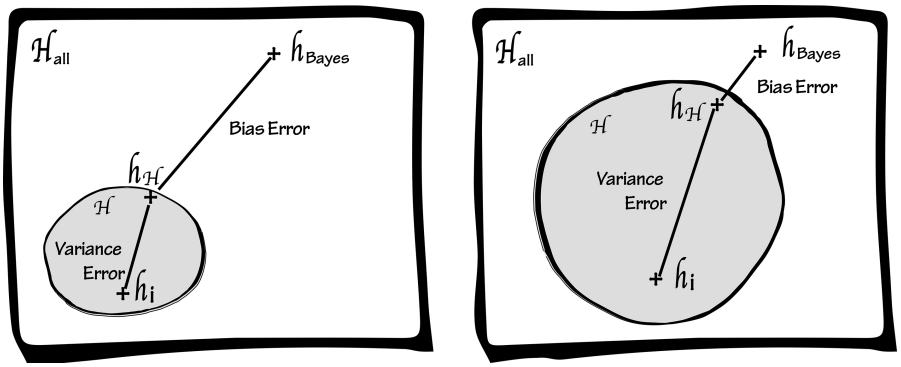


Figure 5.6: Bias and variance errors

kinds of errors due to this constraint:

- **Variance error:** represents how far a classifier h_i is from the best classifier in $\mathcal{H}$, $h_{\mathcal{H}}$. With a strong bias (small hypothesis space), any hypothesis h_i is expected to be closer to $h_{\mathcal{H}}$, there is less variance in the hypothesis space (See figure 5.6). Finding the best hypothesis in a larger hypothesis space is more laborious and, therefore, takes more resources (time and examples) than in a smaller one.

- **Bias error:** represents how far the classifier $h_{\mathcal{H}}$ is from the best classifier h_{Bayes}. With larger, more complex, higher-order, hypothesis spaces we expect $h_{\mathcal{H}}$ to be closer to h_{Bayes} (See figure 5.6).

These two errors compound the generalisation gap, $\Delta(h_i)$:

$$\Delta(h_i) = R(h_i) - R(h_{\text{Bayes}}) \tag{5.10}$$

$$= \underbrace{(R(h_i) - R(h_{\mathcal{H}}))}_{\text{Variance Error}} + \underbrace{(R(h_{\mathcal{H}}) - R(h_{\text{Bayes}}))}_{\text{Bias Error}} \tag{5.11}$$

Machine learning practicioners will recognise here what is called *overfitting* and *underfitting*:

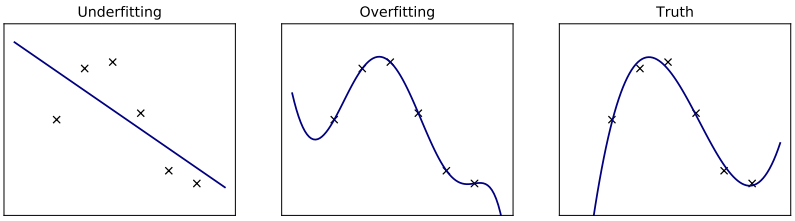


Figure 5.7: Example of underfitting and overfitting in a regression problem.

- **Overfitting:** bias error is small, but variance error is large; High variance is a consequence of fitting to random noise in the training data, rather than the intended outputs.
- **Underfitting:** bias error is large, but variance error is small; The bias error comes from wrong assumptions in the learning algorithm. Strong bias can cause an algorithm to miss relevant relations between inputs and outputs.

It is easy to notice that these two errors are conflicting: the stronger the bias, smaller is the $\mathcal{H} \subset \mathcal{H}_{all}$, smaller is the variance error, but bigger is the bias error; and vice-versa (figure 5.8). This trade-off is the central paradigm of Machine Learning Theory [53], its crucial challenge, and has different names underfitting-overfitting, precision-complexity, and performance-prediction.

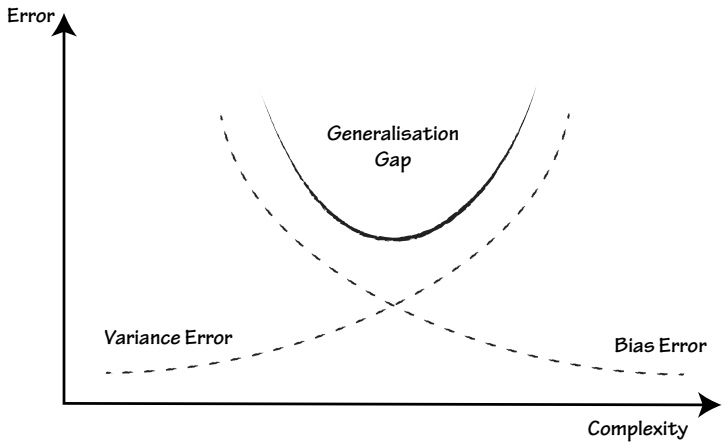


Figure 5.8: Generalisation gap.

The goal of machine learning algorithms is to come up with the simplest model that explains the data, but not simpler.

There are many more complicated explanations possible than simple ones. Therefore, if a simple explanation happens to fit your data, it is much less likely this is happening just by chance.

– Blum [12]

5.4 THE PAC LEARNING MODEL

Up to this point in the chapter, we have described MLT in accordance to Statistical Learning Theory (SLT). Now we will revisit some of what we already explained with the formalism of the PAC model. The PAC model was proposed by Leslie Valiant (see figure 5.9) in 1984 [61]. The fact this work does not cite the previous work of Vapnik and Chervonenkis is an indication that the overlap of CoLT and STL was reinvented. As expected, CoLT look to the learning problem from a computational perspective, while SLT from a statistical one.

"The PAC framework deals with the question of learnability for a concept class $\mathbb{C}$ and not a particular concept" [43]. The PAC model classifies concept classes in terms of their complexities; sample complexity, the number of examples needed, and computation complexity, number of iterations needed, to achieve an approximate solution.

Figure 5.9: Leslie Valiant recieved the Turing Award in 2010.

In the PAC framework, a *concept* c is learnable if there is an algorithm capable of generating, with polynomial time and examples, a general function (the hypothesis h) that with high confidence $(1-\delta)$, has an arbitrarily small error ϵ in any given instance of the problem.

$$\underbrace{\text{Probably}}_{\text{confidence} \geqslant (1-\delta)} \quad \underbrace{\text{Approximately}}_{\text{tolerance} \leqslant \epsilon} \quad \underbrace{\text{Correct}}_{h(\cdot)=c(\cdot)} \tag{5.12}$$

If, with absolute certainty, the hypothesis "imitates" the concept, i. e. there is no error, we can say that there was learning:

$$\exists h \in \mathcal{H} : \; P_{x \sim D}[c(x) \neq h(x)] = 0 \rightarrow \text{learning}. \tag{5.13}$$

But this definition is too restritive. For instance, if $c \notin \mathcal{H}$, there is no way of any h perfectly imitating c. Let us redefine learning with these new relaxed constraint to the absolute error:

$$P_{x \sim D}[c(x) \neq h(x)] = R_D(h) \tag{5.14}$$

$$\exists h \in \mathcal{H} : R_D(h) \leqslant \epsilon, \; 0 < \epsilon < \tfrac{1}{2} \rightarrow \text{learning}. \tag{5.15}$$

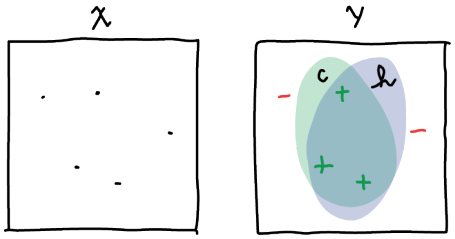


Figure 5.10: The concept versus the hypothesis.

Allowing some tolerance to error, however, is still not sufficient. At one side, a *hypothesis* does not need to be equal to the *concept* to be ***consistent* to the sample**, i. e. to correcly predict every example of the sample. In the figure 5.10, the hypothesis was *lucky*, and there is no difference between the hypothesis and the concept for the particular sample, even though they are different maps of $\mathcal{X}$.

On the other side, it is possible that the sampling:

$$S_n = \{(x_1, y_1), \cdots, (x_n, y_n)\} \sim D^n \tag{5.16}$$

is *unlucky*, and give the same kind of examples for the learning algorithm, an uninformative sample, making it impossible for the hypothesis to *imitate* the concept for all $x \in \mathcal{X}$. In this *unlucky* case, learning would be impossible. Hence, we relax the constraints once more:

$$\begin{aligned} &\exists h \in \mathcal{H},\ 0 < \epsilon < \tfrac{1}{2},\ 0 < \delta < \tfrac{1}{2} : \\ &P_{S\sim D^n}[R_D(h) > \epsilon] < \delta \rightarrow \text{learning}. \end{aligned} \tag{5.17}$$

Nevertheless, if achieving such thresholds demands an unreasonable ammount of data and time, can we really say that learning has happened? What is a reasonable amount of time and examples?

Let d be a number such that representing any vector $x \in \mathcal{X}$ costs at most $\mathcal{O}(d)$ (e. g. $\mathcal{X} = \mathbb{R}^d$), and $\text{size}(c)$ the computational cost of representing a concept $c \in \mathbb{C}$.

Definition 34. *A concept class $\mathbb{C}$ is **PAC-learnable** if there is a learning algorithm $\mathcal{A}$ and a polynomial function $\text{poly}(\cdot,\cdot,\cdot,\cdot,)$ such that for any $0 < \epsilon < \frac{1}{2}$ and any $0 < \delta < \frac{1}{2}$, for any distribution D on $\mathcal{X}$ and for any target concept $c \in \mathbb{C}$, the following holds for any sample size $n \geqslant \text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, d, \text{size}(c))$ [43]:*

$$P_{S\sim D^n}[R_D(h) \leqslant \epsilon] \geqslant (1 - \delta). \tag{5.18}$$

*If $\mathcal{A}$ further runs in $\text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, d, \text{size}(c))$, then $\mathbb{C}$ is said to be **efficiently PAC-learnable**. When such an algorithm $\mathcal{A}$ exists, it is called a **PAC-learning algorithm** for $\mathbb{C}$ [43].*

5.5 PAC BOUNDS

As we stated before, one of the main goals of Machine Learning Theory (MLT) is to guarantee bounds to the error and the number of samples needed (sample complexity) in learning problems. Here we present some of these guarantees as examples on how this theoretical development allow us to make claims on unknown distributions and unseen examples.

5.5.1 Guarantees for finite hypothesis spaces — consistent case

Theorem 8 (Haussler, 1988). *Let $\mathcal{H}$ be a finite hypothesis space, $\mathcal{A}$ a learning algorithm that returns a consistent hypothesis h, i.e. $\hat{R}_S(h) = 0$, for any target concept c and unknown distribution $D = P(X, Y)$.*
Let $|S| = n$, then, $\forall n \geqslant 1$:

$$P[\exists h \in \mathcal{H} : R_D(h) > \epsilon] \leqslant |\mathcal{H}| e^{-\epsilon n} \tag{5.19}$$

Proof. Let $h_{bad}(bad = 1, ..., k)$ be all hypotheses in the space $\mathcal{H}_{bad} \subset \mathcal{H}$ where $\forall h_{bad} \in \mathcal{H}_{bad} : R_D(h_{bad}) > \epsilon$, then:

The chance of a *bad* hypothesis to correctly predict an example is:

$$P_{x_j \sim S}[(c(x_j) \neq h_{bad}(x_j)) = \emptyset] \leqslant (1 - \epsilon) \tag{5.20}$$

$$P_{x_j \sim S}[R_{x_j}(h_{bad}) = 0] \leqslant (1 - \epsilon) \tag{5.21}$$

Therefore, the probability that a *bad* hypothesis will predict correctly all examples in the training sample S_n is:

$$P_{x_1 \sim S}[R_{x_1}(h_{bad}) = 0] \wedge \tag{5.22}$$

$$P_{x_2 \sim S}[R_{x_2}(h_{bad}) = 0] \wedge \tag{5.23}$$

$$\cdots$$

$$P_{x_n \sim S}[R_{x_n}(h_{bad}) = 0] \leqslant \underbrace{(1 - \epsilon) \cdots (1 - \epsilon)}_{n} \tag{5.24}$$

$$P[\hat{R}_S(h) = 0 \wedge R_D(h) > \epsilon)] \leqslant (1 - \epsilon)^n \tag{5.25}$$

We said there are k *bad* hypotheses, then, the probability of any of these *bad* hypothesis predicting correcly all the training sample is:

$$P[h_1 \in \mathcal{H}_{bad} : \hat{R}_S(h_1) = 0 \wedge R_D(h) > \epsilon)] \vee \tag{5.26}$$

$$P[h_2 \in \mathcal{H}_{bad} : \hat{R}_S(h_2) = 0 \wedge R_D(h) > \epsilon)] \vee \tag{5.27}$$

$$\cdots P[h_k \in \mathcal{H}_{bad} : \hat{R}_S(h_k) = 0 \wedge R_D(h) > \epsilon)]$$

$$\leqslant \sum_{1}^{k} (1 - \epsilon)^n \tag{5.28}$$

$$P[\exists h \in H : \hat{R}_S(h_1) = 0 \wedge R_D(h) > \epsilon] \leqslant k(1 - \epsilon)^n \tag{5.29}$$

Finally, as these *bad* hypotheses belong to $\mathcal{H}_{bad} \subset \mathcal{H}$, $k < |\mathcal{H}|$, therefore, we get the theoretical error of h given a precision tolerance of ϵ, and a sample complexity of n examples:

$$P[\exists h \in \mathcal{H} : R_D(h) > \epsilon] \leqslant |\mathcal{H}|(1-\epsilon)^n \tag{5.30}$$

$$(1-x) \leqslant e^{-x}, 0 \leqslant x \leqslant 1 \implies$$

$$P[\exists h \in \mathcal{H} : R_D(h) > \epsilon] \leqslant |\mathcal{H}|e^{-\epsilon n} \qquad \square$$

From the PAC framework:

$$P[\exists h \in \mathcal{H} : R_D(h) > \epsilon] < \delta \tag{5.31}$$

Therefore, Haussler theorem gives us a lower bound on the confidence:

$$\delta > |\mathcal{H}|e^{-\epsilon n} \geqslant P[\exists h \in \mathcal{H} : R_D(h) > \epsilon] \tag{5.32}$$

We can rewrite the Haussler theorem to bound the number of examples needed for learning:

Theorem 9. *A learning algorithm $\mathcal{A}$ can learn a concept c from a class of concepts $\mathbb{C}$ with $n < \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})$ training examples.*

Proof.

$$\delta > |H|e^{-\epsilon n} \qquad \text{(from equation 5.32)}$$

$$e^{-\epsilon n} < \frac{\delta}{|\mathcal{H}|} \tag{5.33}$$

$$-\epsilon n < (\ln\delta - \ln|\mathcal{H}|) \tag{5.34}$$

$$\epsilon n < (\ln|\mathcal{H}| - \ln\delta) \tag{5.35}$$

$$n < \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta}) \tag{5.36}$$

$$n \in \mathcal{O}\left(\frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})\right) \qquad \text{(sample complexity)}$$

$\square$

Strangely, the sample complexity upper bound does not depend on $\mathbb{C}$ or D, but depends logarithmically to the size of $\mathcal{H}$ [23].

5.5.2 No free lunch theorem

IS A UNIVERSAL CONCEPT CLASS LEARNABLE? Let $\mathcal{X} = \{0,1\}^d$, the space of Boolean vectors of size d. A universal concept class $\mathcal{U}_d$ has all subsets of $\mathcal{X}$, i.e. contains all possible classifications for a given instance space $\mathcal{X}$.

$$|\mathcal{U}_d| = 2^{|\mathcal{X}|} = 2^{(2^d)} \tag{5.37}$$

$$|\mathcal{H}| \geqslant |\mathcal{U}_d| \tag{5.38}$$

$$|\mathcal{H}| \geqslant 2^{(2^d)} \tag{5.39}$$

From 9:

$$n \in \mathcal{O}\left(\frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})\right) \tag{5.40}$$

$$n \in \mathcal{O}\left(\frac{1}{\epsilon}(2^d \ln(2) + \ln\frac{1}{\delta})\right) \therefore \tag{5.41}$$

$$n \in \mathcal{O}\left(2^d; \frac{1}{\epsilon}; \ln\frac{1}{\delta}\right) \tag{5.42}$$

Therefore, the sample complexity is not polynomial to d, and $\mathcal{U}_d$ **is not PAC Learnable**. The "no free lunch" theorem [65] states there is no universal concept, therefore, no universal learning algorithm for all tasks. Specifically, averaged over all possible data generating distributions, every classification algorithm achieves the same error when classifying previously unknown points.

5.5.3 Guarantees for finite hypothesis spaces — inconsistent case

Usually, there is no hypothesis in $\mathcal{H}$ consistent with the training sample, due to the stochastic nature of the supervisor or the concept class being more complex than the hypothesis class used by the learning algorithm.

To derive bounds for this inconsistent case, we will use the "law of large numbers".

LAW OF LARGE NUMBERS The law of large numbers states that the mean of random variables ξ_i, drawn i.i.d. from some probability distribution P, converges to the mean of P itself when the sample size goes to infinity.

$$\text{for } n \to \infty,$$
$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \to \mathbb{E}(\xi), \xi_i \sim P. \tag{5.43}$$

Remember: A statistic is a function of random variables that does not depend on parameters.

Based on the fact that a statistic of random variables can be treated as a random variable, we can make the loss function $\ell(x, y, h(x))$ be the random variable ξ from above. From what we can conclude that for a fixed h, the empirical risk converges to the theoretical risk as the sample size goes to infinity:

$$\text{for } n \to \infty,$$
$$\hat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n}(\ell(x_i, y_i, h(x_i)) \to \mathbb{E}(\ell(x, y, h(x)) = R(h). \tag{5.44}$$

CHERNOFF-HOEFFDING INEQUALITY Moreover, we can use the famous *Chernoff-Hoeffding's inequality* to bound the approximation of the risk:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathbb{E}(\xi)\right| \geqslant \epsilon\right) \leqslant 2\exp(-2n\epsilon^2)$$

(Chernoff-Hoeffding's inequality)

$$P(|\hat{R}_S(h) - R(h)| \geqslant \epsilon) \leqslant 2\exp(-2n\epsilon^2) \tag{5.45}$$

Unfortunately, this bound only holds for a fixed function h which does not depend on the training data, but our hypothesis certainly does depend. The reason for such constraint is intuitive. If we let the hypothesis space convey all possible functions and do not restrict our hypothesis to be independent on the training data, we can always generate a function that "memorises" the given sample and has no empirical error. Such function will most certainly not generalise well and invalidate the bound.

Vapnik and Chervonenkis solved this conundrum by using the union bound.

UNION BOUND Even if we are not allowed to select an hypothesis from the space using training data, the bound still holds for any hypothesis that is taken at random. Also, if we enumerate all the functions in $\mathcal{H}$, using the fact that it is finite, the bound still holds for each hypothesis:

$$\begin{aligned}
&P(|\hat{R}_S(h_1) - R(h_1)| > \epsilon) \vee \\
&P(|\hat{R}_S(h_2) - R(h_2)| > \epsilon) \vee \cdots \\
&P(|\hat{R}_S(h_{|\mathcal{H}|}) - R(h_{|\mathcal{H}|})| > \epsilon) \leqslant \sum^{|\mathcal{H}|} 2\exp(-2n\epsilon^2) && (5.46)\\
&\therefore P(\sup_{h\in\mathcal{H}} |\hat{R}_S(h) - R(h)| > \epsilon) \leqslant 2|\mathcal{H}|\exp(-2n\epsilon^2) && (5.47)\\
&P\left[\exists h \in \mathcal{H} : \ |\hat{R}_S(h) - R(h)| > \epsilon\right] \leqslant 2|\mathcal{H}|\exp(-2n\epsilon^2) && (5.48)
\end{aligned}$$

Theorem 10. *Let $\mathcal{H}$ be a finite hypothesis class. Then, for any $0 < \delta < \frac{1}{2}$, with probability at least $1 - \delta$, the following innequality holds [43]:*

$$\begin{aligned}
&\forall h \in \mathcal{H}, R(h) \leqslant \hat{R}_S(h) + \epsilon \\
&R(h) \leqslant \hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}} && (5.49)
\end{aligned}$$

Proof.

$$\begin{aligned}
&P\left[\exists h \in \mathcal{H} : \ |\hat{R}_S(h) - R(h)| > \epsilon\right] < \delta && \text{(from PAC)}\\
&P\left[\exists h \in \mathcal{H} : \ |\hat{R}_S(h) - R(h)| > \epsilon\right] \leqslant 2|\mathcal{H}|\exp(-2n\epsilon^2) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(from equation 5.48)}\\
&\therefore \delta > 2|\mathcal{H}|\exp(-2n\epsilon^2) && (5.50)
\end{aligned}$$

Assuming $\delta = 2|\mathcal{H}|\exp(-2n\epsilon^2)$, we have:

$$\exp(-2n\epsilon^2) = \frac{\delta}{2|\mathcal{H}|} \tag{5.51}$$

$$-2n\epsilon^2 = \ln\delta - \ln 2|\mathcal{H}| \tag{5.52}$$

$$\epsilon^2 = \frac{\ln|\mathcal{H}| + \ln 2 - \ln\delta}{2n} \tag{5.53}$$

$$\therefore \epsilon > 0 \rightarrow \epsilon = +\sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}} \tag{5.54}$$

By definition, $R(h) \geqslant \hat{R}_S(h)$, thus:

$$P\left[\exists h \in \mathcal{H} : \ (R(h) - \hat{R}_S(h)) > \epsilon\right] < \delta \tag{5.55}$$

$$P\left[\forall h \in \mathcal{H} : \ (R(h) - \hat{R}_S(h)) \leqslant \epsilon\right] \geqslant 1 - \delta \tag{5.56}$$

Therefore, with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H}, R(h) \leqslant \hat{R}_S(h) + \epsilon \tag{5.57}$$

$$\forall h \in \mathcal{H}, R(h) \leqslant \hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}} \quad \text{(from equation 5.54)}$$

$$\forall h \in \mathcal{H}, R(h) \leqslant \hat{R}_S(h) + \mathcal{O}\left(\sqrt{\log|\mathcal{H}|}; \sqrt{1/n}; \sqrt{\log 1/\delta}\right) \tag{5.58}$$

□

We can rewrite theorem 10 to bound the sample complexity:

Theorem 11. *A learning algorithm $\mathcal{A}$ can learn a concept c from a class of concepts $\mathbb{C}$ with $n \leqslant \frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2\epsilon^2}$ training examples.*

Proof. From eq. from equation 5.54,

$$\epsilon \leqslant \sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}} \tag{5.59}$$

$$\therefore n \leqslant \frac{\ln|\mathcal{H}| + \ln 2/\delta}{2\epsilon^2} \tag{5.60}$$

□

5.5.4 Guarantees for infinite hypothesis space — inconsistent case

It can be argued that for our use in machine learning, there is no need for guarantees for infinite $\mathcal{H}$ due to the nature of computer hardware and their memory limitations, which already discretise the hypothesis spaces. Therefore, we will just give a general idea of this case.

One of the most striking insights of Vapnik and Chervonenkis is the idea of the *shattering coefficient* (N). Let us take a look at the bound for finite inconsistent case, theorem 10:

$$\forall h \in \mathcal{H},$$
$$R(h) \leqslant \hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}}$$
(finite hypothesis space, inconsistent case)

The $\ln|\mathcal{H}|$ relates to d, the size of the *representation* of the hipothesis space. Another remark worth mentioning is that in the union bound we just added the probabilities of each $h_i \in \mathcal{H}$ without considering where $P(h_j) \cap P(h_k), j \neq k$.

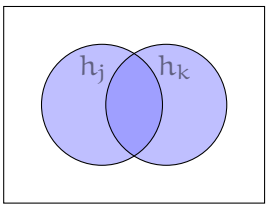


Figure 5.11: $P(h_j) \cap P(h_k)$ is summed twice in the union bound.

In reality, there are several different $h \in \mathcal{H}$ that provide the same map $x \in S \rightarrow y \in \mathcal{Y}$. Therefore, the effective size of $\mathcal{H}$ is smaller than $|\mathcal{H}|$. Using a symmetrization trick [62, section 5.2], Vapnik and Chervonenkis showed that there are at most 2^{2n} effectively different hypothesis. In the PAC framework, $|\mathcal{Y}| = 2$, so if a pattern is a set $\{y_1, \cdots, y_n\}$, there are $|\mathcal{H}| = 2^n$ different patterns, thus, effectively different hypothesis. This number, however, can be even smaller, for example, a certain $y_k, k < n$ can, for example, only accept the value $+$. The shattering coefficient is a growth function, i. e. it measures the number of effectively distinct hypothesis as the size of the sample, n, grows. It is a capacity measure of a hypothesis class. Whenever $N(\mathcal{H}, n) = 2^n$, there exists a sample of size n on which all possible separations of the patterns can be achieved by some $h \in \mathcal{H}$.

We can now rewrite theorem 10 as:

$$\forall h \in \mathcal{H},$$
$$R(h) \leqslant \hat{R}_S(h) + \sqrt{\frac{\ln N(\mathcal{H}, n) + \ln 2/\delta}{2n}} \tag{5.61}$$

Another capacity measure is the famous VC dimension[6].

$$VC(\mathcal{H}) = max\{n \in \mathbb{N} | N(\mathcal{H}, n) = 2^n \text{ for some } S_n\} \tag{5.62}$$

A combinatorial result relates the growth behaviour of the shattering coefficient with the VC dimension:

6 Named after Vapnik and Chervonenkis.

Theorem 12 (Vapnik, Chervonenkis, Sauer, Shelah)**.**

If $\mathrm{VC}(\mathcal{A}) = d$,

$$\forall n \geqslant 1, N(\mathcal{H}, n) \leqslant \sum_{k=0}^{d} \binom{n}{k} \leqslant \left(\frac{en}{d}\right)^{d} \tag{5.63}$$

5.6 CRITIQUES ON MLT

Before a new theory can take place, it is important to understand what went wrong with the current theory. Truth be told: we did not cover current MLT in this chapter which aimed to be an introduction to the subject. There are many topics in active development beyond what was presented here: PAC Bayes, Structural Risk Minimisation, Rademacher complexity, among others.

With this caveat, here we digest some of the critiques on the current state of MLT in two parts, one for general critiques, and another for critiques specific for the case of Deep Learning.

5.6.1 In general

NO ASSUMPTION ON $D = P(X, Y)$ (SEE 5.2.2, ASSUMPTION I.): One of the assumptions of classical learning theory is that "there are no assumptions on $D = P(X, Y)$". Although this assumption means that MLT bounds guarantee approximation to any arbitrary distribution, the ones of practical interest are distributions found in Nature. These practical distributions have some peculiar characteristics that physicists know about [32]: Low polynomial order, locality, symmetry, among others.

ABSENCE OF THE NOTION OF "TIME"(SEE 5.2.2, ASSUMPTION III.): One of MLT assumptions on $P(X, Y)$ is that it is fixed, there is no "time" parameter. Several practical uses of machine learning are in data streams where it is common to have one observation affecting the probability of the future ones [41].

IDENTICALLY INDEPENDENT SAMPLING (SEE 5.2.2, ASSUMPTION IV.): One of the assumptions of Machine Learning is that the datasets are sampled i.i.d. This is assumption is often violated, for example, a machine learning medical application may use data from one hospital to train a model that will be applied all over the world.

The violations are, of course, of practical reasons. But at up to what point can we say that a particular dataset is really i.i.d.? Let us think over the problem of facial recognition. Taking photos at random in an university is not i.i.d, because the people that goes to the universisty is a biased set of the whole population. If we use random images in the

Internet, we may only get the kind of picture people want to display, a bias of intention. Is it not so that if you look harder you can always find some kind of bias in a specific sample? May it be of selection, intention, technical, on this extent?

CONFLICTING OPTIMISATION OBJECTIVES: In the learning theory setting, bias and variance are two conflicting objectives that the learning algorithm tries to minimise. It would be more practical if it was a single sided optmisation problem.

5.6.2 In specific for Deep Learning

VACUOUS BOUNDS Machine Learning Theory cannot explain deep neural networks generalisation performance. Deep learning generalisation gap, according to MLT is in $\mathcal{O}(|\theta| \log |\theta|)$, where $|\theta|$ is the number of parameters of the network [29]. These bounds are vacuous by orders of magnitudes [68].

But deeper and larger networks consistently show better generalisation performance than smaller ones.

INEXPLICABLE PHENOMENA Deep Learning (DL) has several inexplicable phenomena. This does not mean there is no explanation given to these phenomena, but quite contrary there are too many and no clear winning explanation. Here we list some of them.

- Generalisation with more layers: as we explained in this chapter, the current MLT expects models with less parameters to generalise better, that is not what happens in DL. Moreover, Zhang et al. [67] showed that the hypothesis space of DNN is large enough to allow convergence to random labels.
- Disentanglement of semantic factors: the representation of the input in deep layers disentangle semantic factors, i. e. different semantic factors are not strongly correlated in the representation;
- Flat minimum and SGD: Flat minima are points in a function located in a "valley", i. e. the gradient to the neighborhood is small. SGD algorithm running on DNNs seem to have a *preference* for finding minima which are in such "valleys", thus, a preference for *flat minima*. This property of DL is crucial for generalisation. Still, there is no consensus on why this phenomenon occurs.
- Superconvergence: Smith and Topin [54] present that overall trainning time can be shorten and better accuracy achieved by usage of cyclical learning rates. Howard and Ruder [25] propose a small variation of the method, slanted triangular learning rates, and achieve even better performance. This superconver-

gence phenomenon is not well studied and there are only a few conjectures on why it does happen.

- Critical Learning Periods: Achille, Rovere, and Soatto [3] show that "similar to humans and animals, deep artificial neural networks exhibit critical periods during which a temporary stimulus deficit can impair the development of a skill." This finding questions the assumption that the order in which a model experience an evidence does not affect learning.

6 PROPOSAL

Everything is theoretically impossible, until it is done.

– Robert A. Heinlein

This chapter presents the plan for finishing the dissertation.

6.1 THE TORTUOUS PATH SO FAR

Before presenting the plan of work for the future, it will be enlightening to describe our journey so far.

From the beginning, the goal has always been to research how to achieve state of the art results in Deep Learning with significantly smaller datasets. This goal is one of the most significant challenges of the artificial intelligence community; to democratise deep learning, and prevent that it becomes a game of a handful of influential organisations.

The urge for developing methods for learning with less data became even more pressing as our research group shifted its focus to Natural Language Processing (NLP).

TRANSFER LEARNING Thus, we started by researching Transfer Learning (TL). Transfer learning is one of the most critical developments in DL and the reason for much of the current success in the field.

In this research of TL, there was an encouragement to do a literature review, which produced the paper *Research Frontiers in Transfer Learning – a systematic and bibliometric review* [21] (appendix A). In this review, it became clear that:

1. Transfer learning in NLP was a promising field, with exciting new results like ULMFit [25] and BERT [16].

2. There is an outpace between practice and theory in Deep Learning, and several unexplained phenomena and methods that work well in practice without theoretical support.

The premise, then, was that by studying the theory, it would be possible to have insights on developing new methods to achieve more efficiency in transfer learning and, consequently, need less data. And then, we fell down *the rabbit hole*.

MACHINE LEARNING THEORY By trying to understand one of the first papers on the theory of transfer learning (Baxter [10]), it was evident there was a lack of knowledge in Machine Learning Theory. Understanding the current theory became an effort of several months, and culminated in frustration.

The study of MLT explained why the community had turned its back to theory: the current general theory contradicts phenomena experienced in practice and did not guide why methods worked or what is promising.

As we were about to give up the theoretical effort, there was a serendipitous finding: Tishby's video on a new general theory for Deep Learning based on Information Theory [59] started to trend and received positive critiques from widely respected researchers like Geoffrey Hinton and Samy Bengio. And then, another *rabbit hole*.

THE BOTTLENECK THEORY OF DEEP LEARNING The study of the Bottleneck theory of deep learning was arduous for several reasons:

- It assumed knowledge in Information Theory, which was not the case and took other months.
- It is a very new field, and several essential papers were being published as we were doing the literature review.
- It is not well established yet, the work is scattered in several papers and still need to be consolidated.

From the initial literature review, the technical report *An Information Theoretical Transferability Metric* [20] (annexed in appendix chapter B) was produced. Moreover, it was evident that the difficulty found was a great research opportunity.

6.2 PROPOSAL

We propose to consolidate the scattered knowledge on the Information Bottleneck Theory (IBT)in a comprehensive manner.

The literature reviewed so far:

- "Deep learning and the information bottleneck principle" [60] (Tishby and Zaslavsky, 2015)

- "Opening the Black Box of Deep Neural Networks via Information" [51] (Shwartz-Ziv and Tishby, 2017)
- "Emergence of Invariance and Disentangling in Deep Representations" [4] (Achille and Soatto, 2018)
- "Efficient compression in color naming and its evolution" [66] (Zaslavsky et al., 2018)
- *Critical Learning Periods in Deep Neural Networks* [3] (Achille, Rovere, and Soatto 2017)
- "Information Dropout: Learning Optimal Representations Through Noisy Computation" [5] (Achille and Soatto, 2018)
- *Dynamics and Reachability of Learning Tasks* [2] (Achille, Mbeng, and Soatto, 2018)
- *The Information Complexity of Learning Tasks, their Structure and their Distance* [7] (Achille et al., 2019)
- *Representation Compression and Generalization in Deep Neural Networks* [52] (Shwartz-Ziv and Tishby, 2019)
- "On the information bottleneck theory of deep learning" [46] (Saxe et al., 2019)
- *Where is the Information in a Deep Neural Network?* [6] (Achille and Soatto, 2019)
- "Emergent Properties of Deep Neural Networks" [1] (Achille, 2019)

6.3 SCHEDULE

IBT SOTA UPDATE Besides the literature reviewed so far, we will keep checking the state-of-the-art in the Information Bottleneck Theory. This is a background task that may take several months.

USING IBT INSIGHTS IN RESEARCH The study of the Information Bottleneck Theory provided some insights for new research: the relationship of superconvergence and the phase transition between information conversion phase and information forgetting phase; the difference in complexity measures between transfer learning and multi-task learning; the evidence of the bottleneck limit in natural languages and its application in automatic language translation; etc. We expect that the research in this area allow us to publish at least one paper in a reputable conference.

SEMINARS ON IBT As preparation for the dissertation writing, and also as part of our research group effort in knowledge dissimination, seminars on the Information Bottleneck Theory will be presented and open to other students.

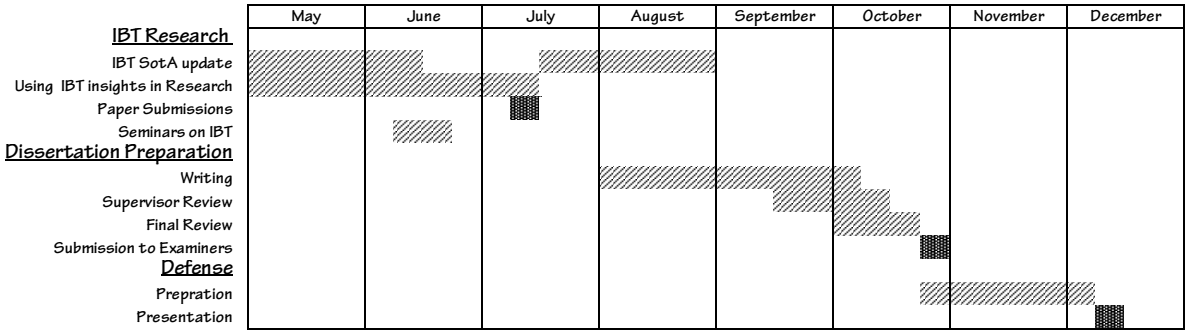


Figure 6.1: Work plan for the rest of the dissertation.

6.4 FINAL CONSIDERATIONS

To the extent of our knowledge, this is the first work that consolidates scattered knowledge of this brand new and promising general theory of deep learning in a comprehensive manner.

Its most prominent research contribution is to guide other researchers so that they do not need to take the same tortuous path we took, and facilitate their journey in this brand new promising subfield.

Some of the challenges are to keep pace with the fast development of publications, at the same time, produce original articles and define a clear scope for the writing of the dissertion avoiding new *rabbit holes*.

Part I

APPENDIX

A RESEARCH FRONTIERS IN TRANSFER LEARNING: A SYSTEMATIC REVIEW

Artificial Intelligence Review manuscript No.
(will be inserted by the editor)

Research Frontiers in Transfer Learning

a systematic review

Frederico Guth · **Teófilo Emidio de Campos**

Received: date / Accepted: date

Abstract Humans can learn from very few samples, demonstrating an outstanding generalization ability that learning algorithms are still far from reaching. Currently, the most successful models demand enormous amounts of well-labeled data, which are expensive and difficult to obtain, becoming one of the biggest obstacles to the use of machine learning in practice. This scenario shows the massive potential for Transfer Learning, which aims to harness previously acquired knowledge to the learning of new tasks more effectively and efficiently. In this systematic review, we apply a quantitative method to select the main contributions to the field and make use of bibliographic coupling metrics to identify research frontiers. We further analyze the linguistic variation between the "classics" of the field and the "frontier" and map promising research directions.

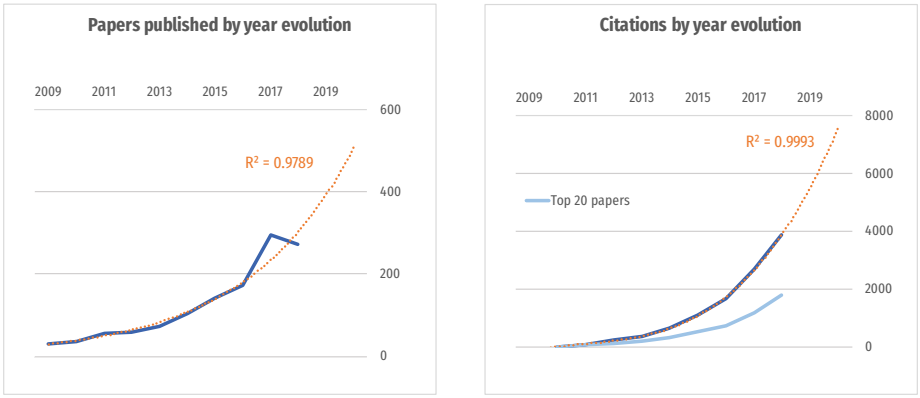
Keywords transfer learning, systematic review, bibliographic analysis

1 Introduction

Currently, machine learning algorithms can recognize objects, people, and places at a superhuman accuracy (L. Fei-Fei, 2017); they can diagnose skin cancer better than dermatologists (Guth and de Campos, 2018), and can even see through walls using radio signal analysis (Zhao et al., 2018). Despite all that, most successful models demand astronomical amounts of well-labeled data that are expensive and difficult to gather. That is because the standard model training procedure starts *tabula rasa*, i.e., with random initialization of model parameters (Ruder, 2019).

Learning this way, from a blank state, is contrary to the way humans do. Every day, we transfer knowledge: knowing how to play the piano makes it easier to learn pipe organ; knowing Portuguese makes it easier to learn Spanish. People use previously obtained knowledge to more effectively and efficiently learn new things (Pan and Yang, 2010). Learning algorithms are still far from reaching this outstanding generalization capability. Recent studies (Jia and Liang, 2017) show that current algorithms hardly generalize further than the data seen during training.

Fred Guth and Teo de Campos
Departamento de Ciência da Computação
Universidade de Brasília
70.910-900 - Brasília -DF
Brazil
E-mail: fredguth@fredguth.com (Fred Guth) E-mail: teodecampos@unb.br (Teo de Campos)



(a) (b)

Fig. 1: Evolution of research production (a), and citations (b), on *Transfer Learning* in the last 10 years, and exponential growth projection with high coefficient of determination (R^2), despite publication decay in 2018. In (b) one can also see that the top 20 most cited papers are responsible for almost half the citations. (Data source: Web of Science, march/2019)

This context reveals the huge unreached potential of Transfer Learning (TL), which aims to leverage prior experience to efficiently and effectively learn new tasks. In practice, TL tends to be applied on an *ad hoc* basis, where the transfer methods are simple extensions of the learning algorithms used (Torrey and Shavlik, 2010). Such importance with a lack of consolidated methods and theories indicates this is a promising field for research. As Andrew Ng (Ng, 2016) says: "Transfer Learning will be the next driver of Machine Learning success across industries". From this perspective, it is understandable the growing interest on the subject (Figure 1).

1.1 Objectives

Our research questions are:

Q1. What are the research frontiers in Transfer Learning?
Q2. Is it possible to base this evaluation on bibliometrics?

To answer these questions, we will first review the literature and reveal the main contributions to the field and how they relate to each other.

1.2 Contributions

C1. We present an updated systematic review of the literature on Transfer Learning, using the TEMAC framework (Section 2). A method that helps us focus on high impact contributions.
C2. We extended TEMAC to analyze linguistic variations of abstracts using ScatterText (Kessler, 2017), which, up to our knowledge, is an original usage of this visualization tool.
C3. We identify, with bibliometric analysis support, the directions of research frontiers, and the open problems of the field.

1.3 Overview

In this brief introduction, we will present *Related Works*. In the next section (section 2), we will explain our research method and the quantitative analysis to support our findings. In section 3, the results, which are indeed the *Literature Review*, are shown. *Open Problems* are discussed in section 4. Finally, we conclude in section 5, presenting answers for our research questions.

1.4 Related Work

As we specify in section 3, there are already literature reviews in the field of *Transfer Learning*, and they tend to be well cited. Noteworthily, Pan and Yang (2010)'s survey is the most cited paper in the field, and Hohman et al. (2018) systematically reviews the literature on visual analytics in *deep learning* with the gripping method of five W's and one H (Why, Who, What, How, When, and Where). Still, to the best of our knowledge, our literature review is the first in *Transfer Learning* to be based on a bibliometric method. Similar to what we propose in our field, Park and Yoon (2018) applies a meta-analytical approach to identify research frontiers in *Pattern Recognition*, and Bhattacharya (1997) uses bibliometrics to identify research frontiers in *Physics*.

2 Method: Literature review with a quantitative approach

Our literature review uses the bibliometric approach of Mariano and Rocha (2017) (TEMAC), which aims to give quantitative support to literature selection.

TEMAC consists of:

a) research preparation;
b) data presentation and interrelationships;
c) detailing, synthesis and validation.

2.1 Research Preparation

On 31st of March, a search on *Clarivate Analytics Web of Science (WoS)* shown in Figure 2 resulted in 1,289 articles found. It is noticeable that the interest in the subject is thriving (Figure 1), and it is possible to project that in three years the number of articles will double (exponential growth, coefficient of determination: $R^2 = 0,9789$).

A restricted search with only recent works (Figure 3) was made for the section 3.5 (*Research Frontiers*).

2.2 Data presentation and interrelationship

In this stage we analyse (see §3.3):

1. most cited articles (Figure 5);
2. number of articles evolution year by year(Figure 1(a));
3. citations evolution year by year (Figure 1(b));

Results: 1,289 (from Web of Science Core Collection)

```
You searched for:
TOPIC: (‘‘transfer learning’’)
Refined by:
WEB OF SCIENCE CATEGORIES:
  ( COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE )
AND LANGUAGES: ( ENGLISH )
AND RESEARCH AREAS: ( COMPUTER SCIENCE )
Timespan: 2009-2019.
Indexes: SCI-EXPANDED, CPCI-S, ESCI.
```

Fig. 2: *“10yearsSearch”*: Search parameters on *Web of Science*.

Results: 384 (from Web of Science Core Collection)

```
You searched for:
TOPIC: (‘‘transfer learning’’)
Refined by:
WEB OF SCIENCE CATEGORIES:
( COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE )
AND LANGUAGES: ( ENGLISH )
AND RESEARCH AREAS: ( COMPUTER SCIENCE )
AND PUBLICATION YEARS: ( 2019 OR 2018 OR 2017 )
AND DOCUMENT TYPES: ( PROCEEDINGS PAPER )
Timespan: 2009-2019.
Indexes: SCI-EXPANDED, CPCI-S, ESCI.
```

Fig. 3: *“3yearsSearch”*: Search parameters for frontier analysis.

4. most published and cited authors (Figure 5);
5. most published and cited conferences (Figure 5);
6. most published and cited institutions (Figure 5);
7. countries by research production (Figure 5);
8. keyword frequency (Figures 7 and 8).

2.3 Synthesis and Validation

a) **Co-citation analysis**: Co-citation measures the frequency in which two papers are cited in the same reference list, and it is assumed that they are “pieces” of the same “knowledge structure”. Co-citation analysis, therefore, maps the intellectual inheritance of a research field by identifying impactful works, but neglects the research frontier as those works had less time to be cited (Vogel and Güttel, 2012).
The free software VOSviewer (van Eck and Waltman, 2009) was used to cluster works cited by the *“10yearsSearch”* selected articles. With that, three knowledge clusters where identified (Figure 4).

b) **bi-coupling analysis**: Bibliographic coupling occurs when two papers have at least one reference in common. Papers are, thus, said to be coupled if their references overlap (Vogel and Güttel, 2012). As it is possible to state a chronological order among cited and citing works, bibliographic coupling allows us to map research “generations” and,

therefore, identify which research is in the leaves of this tree, i.e., are in the frontier of the field. It is important to note that, in this context, being in the frontier is just a chronological incident and does not mean it is a promising work. This limiting aspect of the quantitative approach points to the need for a qualitative complement to identify the "classics of the future".

In the TEMAC framework, the bibliographic coupling should be done in a period not longer than the last three years. In our analysis, we restricted the period from 2017 to March 2019 and the only selected works from conference proceedings (Figure 3), assuming that these works have a shorter period of review and publication and, therefore, represent what is newer in the field.

c) **Textual analysis (*bag-of-words* with tf-idf)**:in this analysis articles are viewed as *bag-of-words* and the concept of **tf-idf** is used to define which words better identify each paper. For instance, which words better explain research in frontier *vis-à-vis* those which explain articles from *"10yearsSearch"* (see Figure 8). The metric **tf-idf** is defined as:

$$\text{tfidf}(t,d,D) = \text{tf}(t,d) \cdot \text{idf}(t,D) \tag{1}$$

where $\text{tf}(t,d)$ is the term frequency t in document d and $\text{idf}(t,D)$ is the inverse of the frequency of t in the document set D (*corpus*).

$$\text{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}} \tag{2}$$

$$\text{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{3}$$

where:

N: size of corpus in number of documents $N = |D|$

$|\{d \in D : t \in d\}|$: number of documents where term t appears (i.e., $\text{tf}(t,d) \neq 0$). To avoid division by zero if the term is not in the corpus, the denominator is adjusted to: $1 + |\{d \in D : t \in d\}|$.

The metric **tf-idf** is the basis of the visualization tool *ScatterText* (Kessler, 2017) which was used to generate the Figures 6 and 8.

3 Literature Review

3.1 A brief history of Transfer Learning

Since 1995, when a NIPS[1] workshop on "Learning to Learn" discussed the need for machine learning to retain and reuse previously acquired knowledge, research on transfer learning, although sometimes called by different names (*learning to learn, life-long learning, knowledge transfer*) has attracted more and more attention (Pan and Yang, 2010) (see Figure 1).

In 2005, a DARPA project announcement used, maybe for the first time, the term *transfer learning*, defined as the goal of extracting knowledge from one or more *source tasks* and applying it to *target tasks* (Pan and Yang, 2010). A search on Web of Science can confirm that the first articles to use the term "transfer learning" appear in 2005.

In 2012, a deep neural network used by Alex Krizhevsky and team in the ImageNet Challenge (ILSVRC) was 41% better than the second place, an outstanding result that ignited the

[1] Currently, the Neural Information Process Systems conference is called NeurIPS.

exponential growth on deep learning research. Such success highlighted the importance of data availability for the advancement of artificial intelligence. It gave birth to a new era in transfer learning. Despite the cost of learning with big datasets like ImageNet, trained models proved to be easily suitable to initialize models for different tasks (Ruder2019Neural, donahue2014decaf). This "fine-tunning" approach allows good results on many tasks with orders of magnitude less data (see Section 3.4.3).

In the present moment, *transfer learning* is a customary topic in prestigious conferences like CVPR, ICCV, ICPR e NeurIPS (see Figure 5, *Top 10 conferences*).

3.2 Notation and Definitions

Transfer Learning entails the concepts of domain and task. According to the notation of Pan and Yang (2010), a domain $\mathscr{D}$ is composed by a feature space $\mathscr{X} \subset R^d$ and a marginal distribution $P(\mathrm{X})$, from where samples $\mathrm{S} = \{x_1, \cdots, x_n\} \in \mathscr{X}$ are drawn. In an image classification problem, for instance, $\mathscr{X}$ is the space of all possible images with a certain dimension and number of channels, x_i is an image, and S is the training *dataset*.

Given a domain $\mathscr{D} = \{\mathscr{X}, P(X)\}$, a task $\mathscr{T}$ can be statistically defined by the conditional distribution $P(\mathrm{Y}|\mathrm{X})$, that is, $\mathscr{T} = \{\mathscr{Y}, f(\cdot)\}$, where $f(\cdot)$ is a goal function that given $x_i \in \mathscr{D}$, predicts its corresponding $y_i \in \mathscr{Y}$.

Be $\mathscr{D}_S$ the source domain and $\mathscr{T}_S$ the source task, $\mathscr{D}_T$ the target domain and $\mathscr{T}_T$ the target task, **transfer learning** aims to help learning the function $f_T(\cdot)$ in $\mathscr{D}_T$ using knowledge from $\mathscr{D}_S$ and $\mathscr{T}_S$, where $\mathscr{D}_S \neq \mathscr{D}_T$ or $\mathscr{T}_S \neq \mathscr{T}_T$.

3.3 An overview of transfer learning research

Pan, S. is the most cited author (see Figure 5, Top 10 authors) with 2706 citations. Such impact is mainly due to *A Survey on Transfer Learning* (Pan and Yang, 2010), which is the most cited article in the field with 2240 citations. The main contributions of this article are to present definitions, notations, and taxonomy for transfer learning that made sense for the research community. It was published in IEEE, a journal with an impact factor of 2,775, and that ranks at the 33rd position on InCite JCR for the *Computer Science, Artificial Intelligence* category, which means it is not a usual publication for *transfer learning* research.

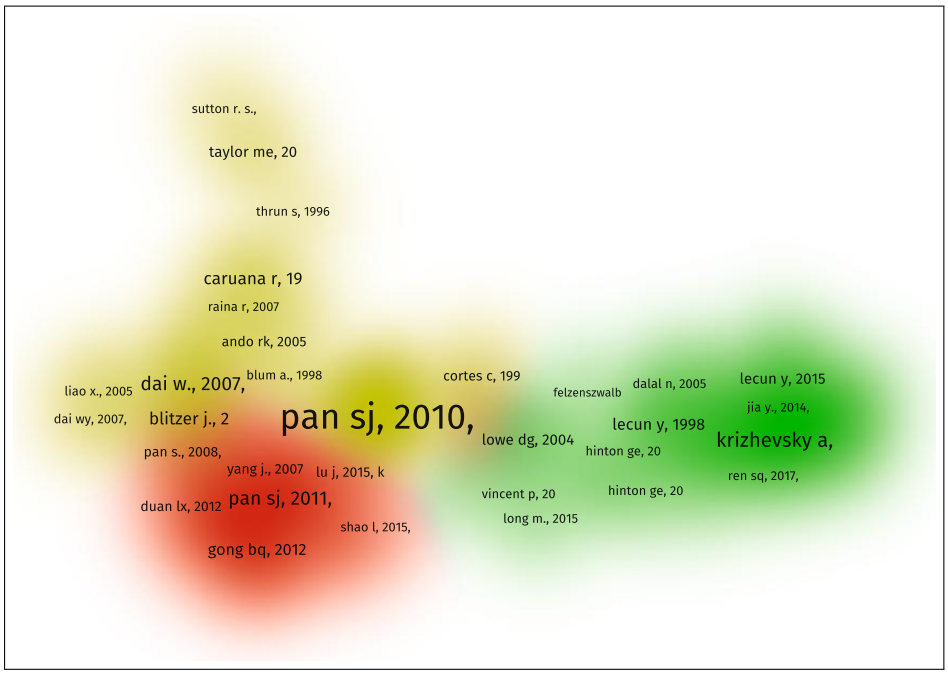
China is the most productive country in the field, followed by the United States and the UK.

Most articles are published in conference proceedings, 63%. CVPR[2] is the conference with most articles on *Transfer Learning*, but ECCV[3], despite ranking only 4$^{\text{th}}$ in production, is the most cited conference. It is also notable that computer vision conferences have been the most popular venue for TL research. It is worth mentioning the lack of conferences with NLP focus in this list and that among the 20 most cited articles, none is on language.

3.4 The classics

Co-citation analysis using *VosViewer* shows 3 knowledge clusters. They present a strong temporal component and can be seen as waves: the first encompasses publications prior to 2011 (the mode is 2006), shown in yellow in Figure 4; the second goes from 2011 to 2014

[2] Conference on Computer Vision and Pattern Recognition [3] European Conference on Computer Vision



Data source: Web of Science (march/2019). Tool:VosViewer(van Eck and Waltman (2009)

Fig. 4: Knowledge clusters from the co-citation analysis. Clusters represent papers that are normally cited together in the list of references of the 10yearsSearch articles.

(mode 2012) and is in red; and the third and last wave, green in the figure, includes articles from 2012 to present day (mode 2014).

3.4.1 First wave

One of the main characteristics of this wave is the strength of its theoretical component. Some of its works present whole classes of transfer learning: Thrun (1996) and Caruana (1997) introduce *Multi-Task Learning* and the idea that auxiliary tasks introduce inductive bias that helps the learning convergence; Chapelle et al. (2006)'s book on *Semi-Supervised Learning*; Raina et al. (2007) on *Self-taught learning*; Vapnik (1998) and the fundamentals of statistical learning theory.

Another component of this wave is the Domain Adaptation articles that aim to learn latent characteristics of reduced dimensionality, assuming that in the latent space source and target domains are similar: Ando and Zhang (2005), Blitzer et al. (2006) and III and Marcu (2006), for instance.

Lastly, some of the most cited articles of this wave are surveys: Pan and Yang (2010), Taylor and Stone (2009); not by coincidence they were published by the end of the wave, organizing the research material up to that point in time.

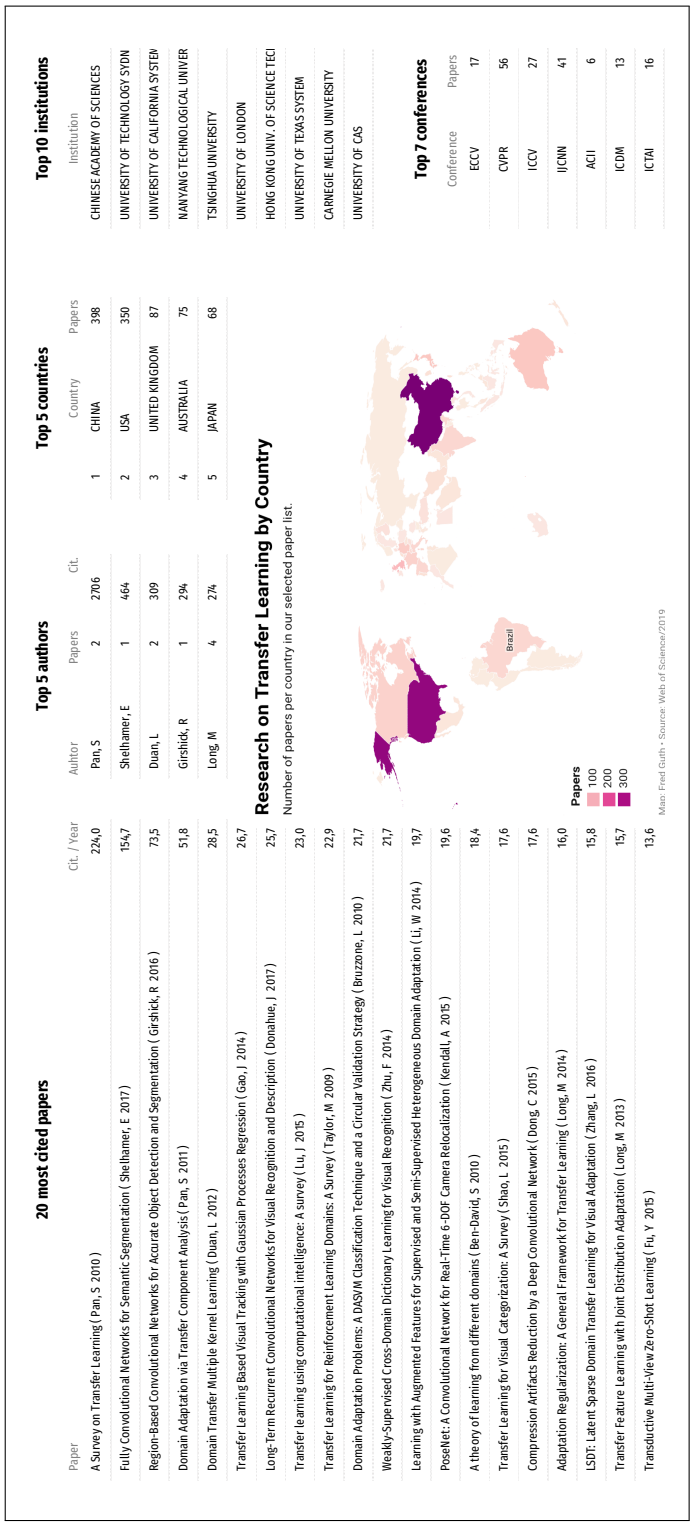


Fig. 5: Overview of transfer learning research.

3.4.2 Second wave

Articles on Domain Adaptation dominated the second wave. Some kept the focus on latent characteristics like Pan et al. (2011), but most articles try to select samples from source domain that are similar to the target domain: Ben-David et al. (2009) approach distribution similarities of domains distributions from a theoretical standpoint; Si et al. (2010) tries to minimize divergence among domains; Common to this trend of work are approaches based on sample classifiers, mostly *Support Vector Machines* (SVMs): Yang et al. (2007), Bruzzone and Marconcini (2010), Duan et al. (2012), among others.

We also find works on unsupervised domain adaptation: Gong et al. (2012), Fernando et al. (2013).

It is noteworthy that among the articles in our search, there was no *survey* "closing" this wave. An excellent example of this kind of work would be the chapter *A Comprehensive Survey on Domain Adaptation for Visual Applications* (Csurka, 2017). Unfortunately, this work is not indexed by *WoS* and, therefore, could not be found in our search results.

3.4.3 Third wave

The third wave cluster includes transfer learning approaches in *deep learning* context, which we can call *deep transfer learning*.

Here we find classical deep learning literature like: Hinton et al. (2006), maybe the seminal article on *Deep Learning*; Bengio (2009) presents *representation learning* and curriculum, a list of tasks to be learned in sequence, because they have growing levels of complexity; LeCun et al. (2015)'s review on *Deep Learning* for *Nature*. Here we also find Krizhevsky et al. (2012)'s article on *AlexNet*, which by winning the 2012 ImageNet challenge (ILSVRC) with a margin of almost 40% over the second place, gave birth to the current "gold rush" of *Deep Learning*.

Some articles in this wave are about large datasets, an essential component of deep learning success:

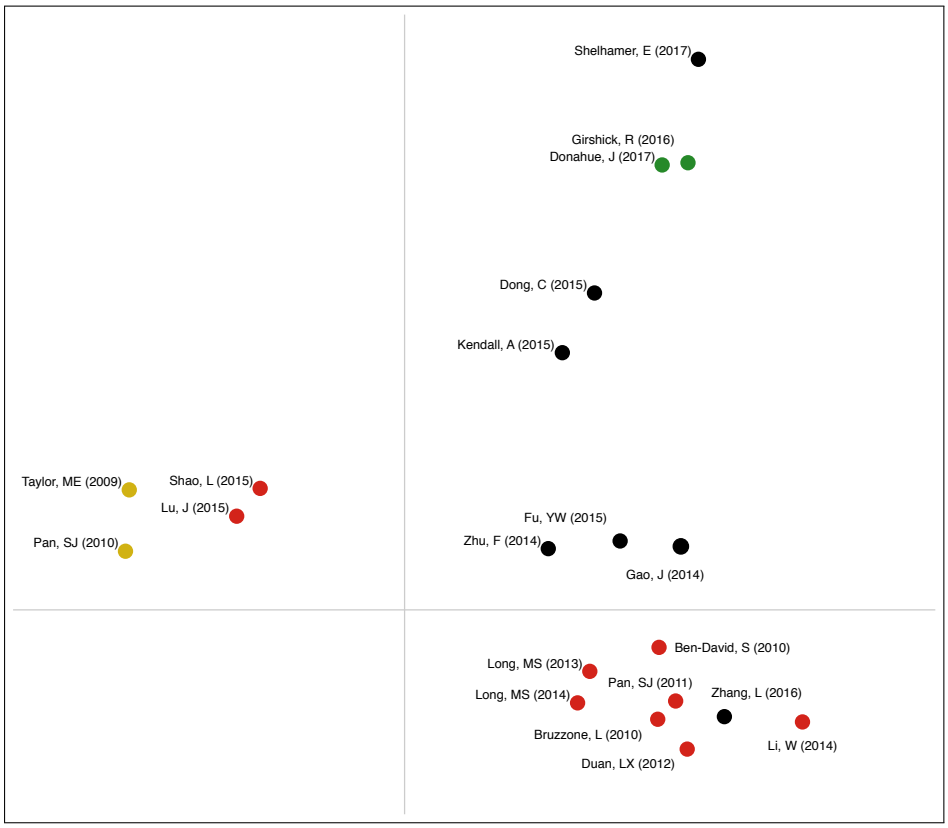
Deng et al. (2009) presents ImageNet and suggest that models learned on ImageNet can be used to more efficiently learn new domains;

Everingham et al. (2009) is about Pascal VOC; Russakovsky et al. (2015) analyses ImageNet's impact on different computer vision problems and, therefore, the role of transfer learning.

There are also articles that present model architectures that use pre-trained models for feature extraction or can be used after training as such: Girshick et al. (2014), which presents the R-CNN model, He et al. (2016), ResNet, and Simonyan and Zisserman (2014), object detection; Fan et al. (2010) human tracking; Long et al. (2015) on semantic segmentation; among others.

Transfer Learning became so ubiquitous in *Deep Learning* that Mahajan et al. (2018) claims that *not* pre-training models with ImageNet in computer vision problems is now considered foolhardy.

For this reason, it is quite difficult to point out articles with the sole focus on transfer learning. Some worth mentioning are: Glorot et al. (2011) which proposes using *deep neural networks* to learn common representation among domains; Donahue et al. (2014) and Oquab et al. (2014) which give some theoretical support to the fine-tuning approach.



Data source: WoS (março/2019). Tool:ScatterText (Kessler (2017)

Fig. 6: Top 20 most cited articles visualized by proximity of *bag of words*.

3.4.4 Textual analysis results

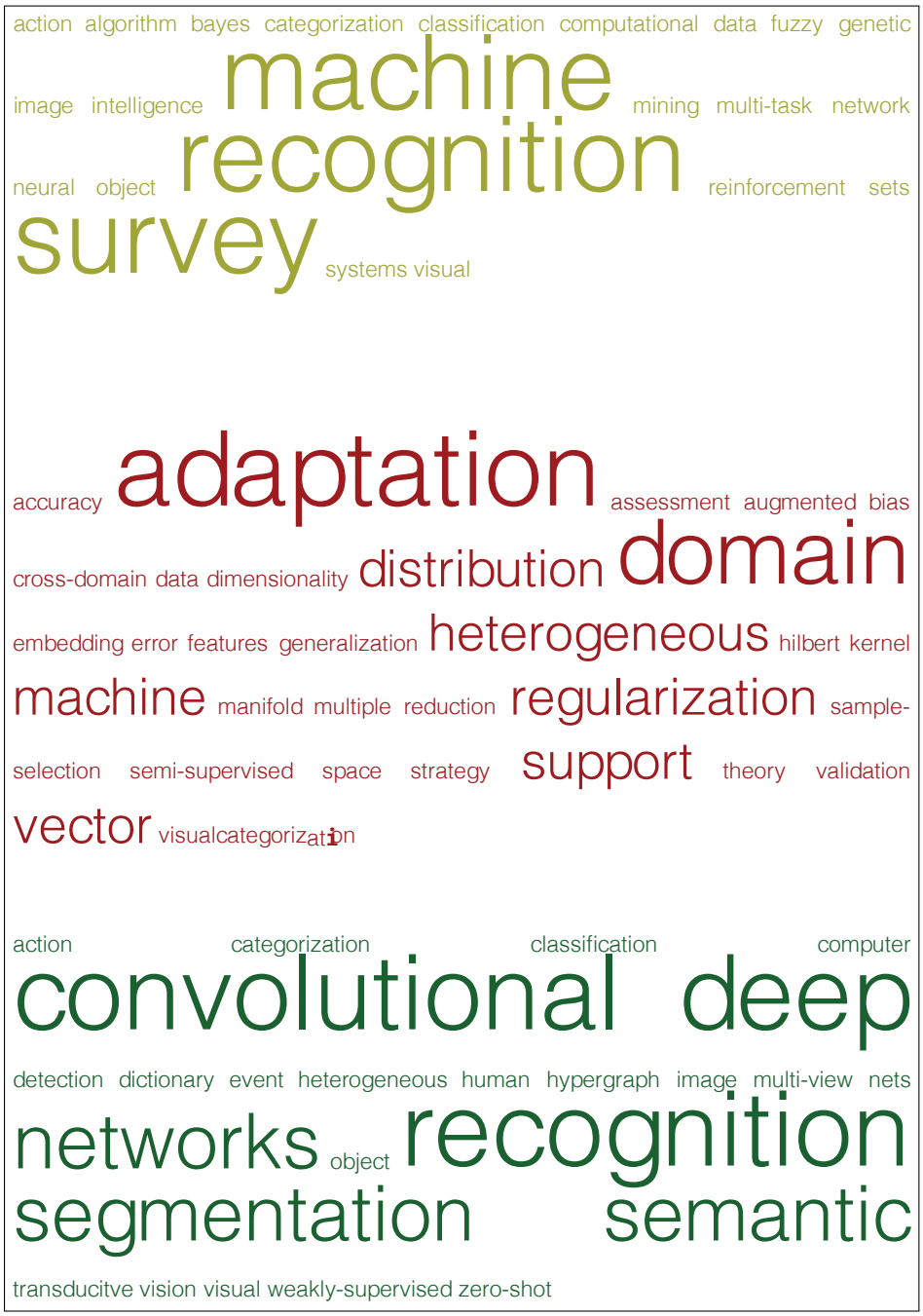
The 20 most cited articles represent almost half the citations (Figure 1b). Consequently, we assume that a textual analysis of this subset of *"10yearsSearch"* is a good proxy for the whole.

By a different method, we approximately cluster in the same way as before. In the Figure 6, each quadrant represents a cluster in Figure 4. That is a strong validation of this analysis.

In Figure 7 it is possible to see the "word cloud" for each quadrant of chart in Figure 6. Those terms correspond well with the qualitative analysis of Section §3.4.

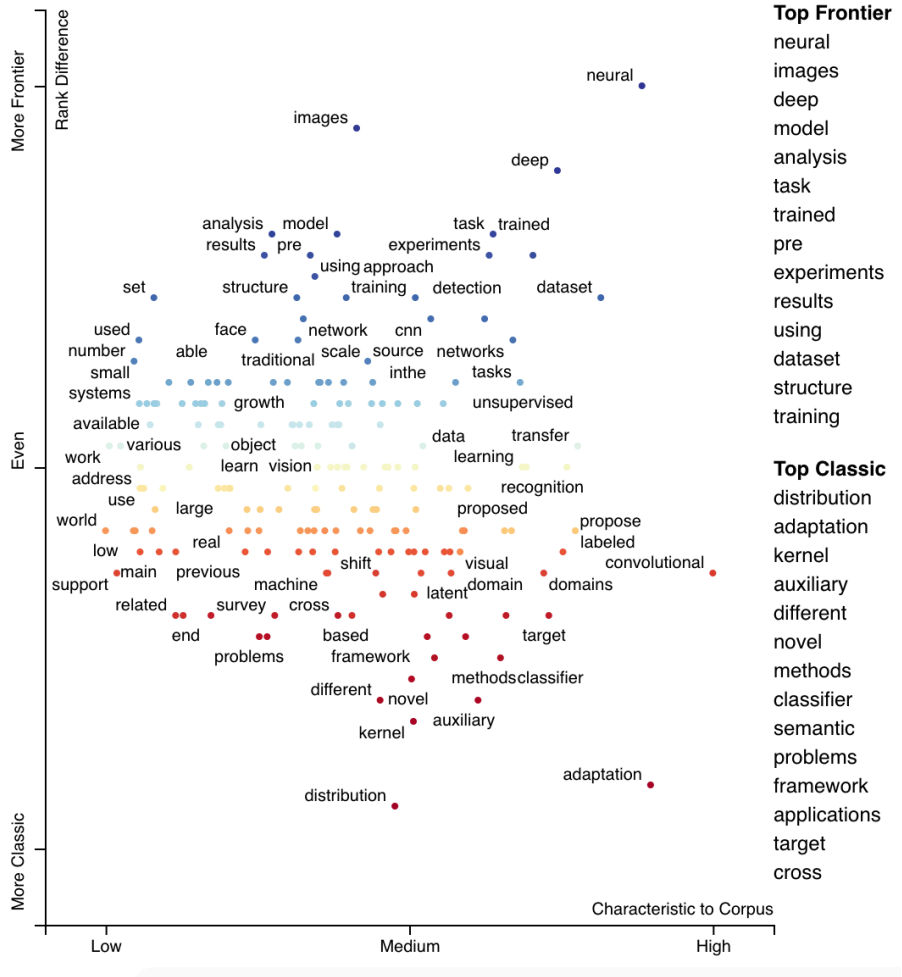
3.5 Research frontiers

To identify the research frontiers in transfer learning, two analysis were performed:



Data source: WoS (march/2019). Tool:TagCrowd

Fig. 7: Word clouds for all quadrants of chart 6.



Data source: *WoS*(march/2019). Tool:ScatterText(Kessler (2017)

Fig. 8: Visual analysis of "frontier" terms versus "classic" terms in the *Transfer Learning* context.

3.5.1 Textual Analysis

We used *ScatterText* (Kessler, 2017) to visualize the terms which better represent *"3yearsSearch"* versus the ones that better represent *"10yearsSearch"*, resulting in Figure 8.

The terms more related with the research frontier are *deep*, *neural*, *images* and words like *cnn*, *trained networks* and *datasets*. That matches our analysis of the third wave in §3.4.3. The terms more related with *"10yearsSearch"* are: *distribution*, *domain*, *adaptation*, *auxiliary* (from auxiliary tasks), *kernel*; which are quite similar to what we found in second wave §3.4.2.

3.5.2 Bibliographical Coupling

The result of the bibliographical coupling analysis using VosViewer (van Eck and Waltman, 2009) with *"3yearsSearch"* articles can be seen in Figure 9.

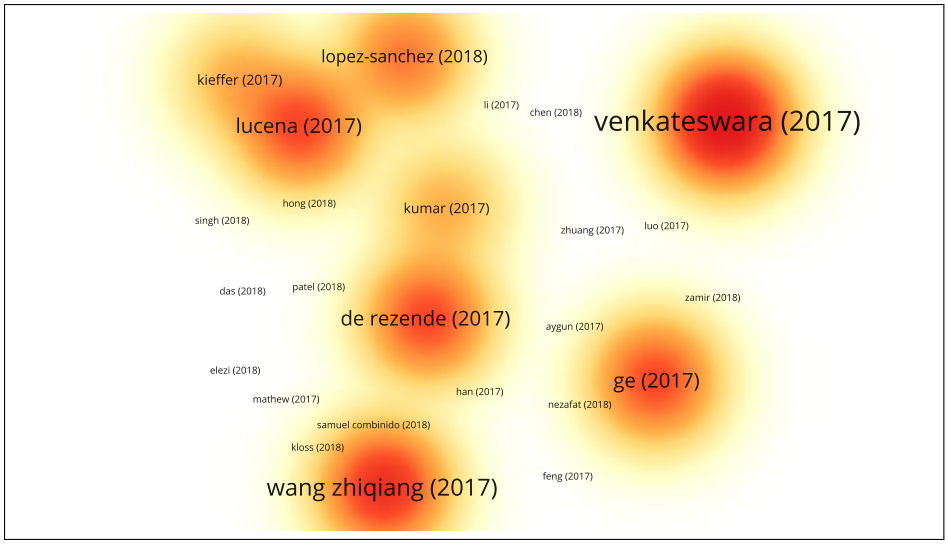
Among the clusters, we have: Lucena et al. (2017) and Rezende et al. (2017), both applications of *deep convolutional networks* for new problems, *face anti-spoofing* and *fake images detection*, respectively. Venkateswara et al. (2017) and Ge and Yu (2017) propose new Domain Adaptation methods using *Deep Learning*. Clearly every work in this frontier is somehow about *Deep Transfer Learning*.

3.5.3 The classics of tomorrow

As stated previously, the quantitative analysis using bibliographical coupling can show which articles are in the frontier (the ones on Deep Transfer Learning) but not which of them are promising. For such, a qualitative analysis of candidate articles is crucial.

In this context, we here highlight some works that we believe have the potential to become classics of tomorrow:

- Zamir et al. (2018): one of the best papers of CVPR 2018, this paper explores relations between tasks, measuring them, and building a graph that can be used to define sequences of training that lead to smaller sample complexity for a specific task.
- Howard and Ruder (2018): by exploiting transfer learning, the ULMFit model was able to perform between 18 and 24% better than the previous state of the art for some NLP problems and can become to language what Krizhevsky et al. (2012) was for computer vision.



Data source: *WoS*(march/2019). Tool:VosViewer

Fig. 9: Heatmap of the bibliographical coupling analysis.

- Zhu et al. (2017): propose *Generative Adversarial Networks (GANs)* for unsupervised domain transfer. We are extremely optimistic regarding the potential of GANs in transfer learning.
- Ruder (2019): presents a new taxonomy for transfer learning in the context of NLP.

4 Open Problems

The overview of a systematic review allows us to perceive some gaps on the field accumulated knowledge. Some open problems are:

1. Metrics: no specific metrics for transfer learning.
2. Taxonomy: current taxonomy (Pan and Yang, 2010) focuses too much on domain adaptation and too little on inductive transfer learning. Also, we need to include more novel ideas like GANs and *autoencoders*.
3. NLP: there is still too little about transfer learning for NLP, with the exception made for Ruder (2019).
4. Theory: on the first wave, the role of theory was trying to indicate promising approaches. Today, some approaches work well in practice, but the theory does not explain why. It is important to know why.

5 Conclusion

In this systematic literature review, it was shown that *deep transfer learning* is in the research frontier of transfer learning. It is a very broad sub-field and encompasses a) usage of pre-trained models as feature extraction for new applications; b) finding latent representations among domains; c) unsupervised style transfer between domains; among others. In this work, we were guided by the TEMAC method and extended it with other analytical tools to validate our conclusions. By doing so, it was possible to demonstrate that it is possible to identify research frontiers with a quantitative-based analysis on bibliographical data, which answers our research questions.

In future work, some important matters need to be addressed: First, Pan and Yang (2010) taxonomy is outdated: for obvious reasons, it focuses on the kind of work that had been done up to a decade ago and does not give much guidance for classifying articles in the research frontier. Besides, the systematic review method deserves some improvement: 1) It could include other bases as *Scopus* and *Google Scholar*; 2) We could expand the research queries to include other terms: *multi-task learning*, *domain adaptation*, and even some that are not as used anymore like *learning to learn* and *lifelong learning*; 3) it could encompass the summarization of most important articles, preferably adopting a framework like the five W's and one H (Hohman et al., 2018).

Lastly, from this review it became clear to us the need to improve the theory of Deep Learning to better explain how and why the transfer of knowledge happens in deep neural networks.

References

Ando RK, Zhang T (2005) A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6(Nov):1817–1853

Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2009) A theory of learning from different domains. Machine Learning 79(1-2):151–175, DOI 10.1007/s10994-009-5152-4, URL https://doi.org/10.1007/s10994-009-5152-4

Bengio Y (2009) Learning deep architectures for AI. Foundations and Trends® in Machine Learning 2(1):1–127, DOI 10.1561/2200000006, URL https://doi.org/10.1561/2200000006

Bhattacharya S (1997) Cross-national comparison of frontier areas of research in physics using bibliometric indicators. Scientometrics 40(3):385–405

Blitzer J, McDonald R, Pereira F (2006) Domain adaptation with structural correspondence learning. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '06, pp 120–128, URL http://dl.acm.org/citation.cfm?id=1610075.1610094

Bruzzone L, Marconcini M (2010) Domain adaptation problems: A DASVM classification technique and a circular validation strategy. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(5):770–787, DOI 10.1109/tpami.2009.57, URL https://doi.org/10.1109/tpami.2009.57

Caruana R (1997) Multitask learning. Machine learning 28(1):41–75, DOI 10.1023/a:1007379606734, URL https://doi.org/10.1023/a:1007379606734

Chapelle O, Scholkopf B, Zien A (eds) (2006) Semi-Supervised Learning. The MIT Press, URL https://doi.org/10.7551/mitpress/9780262033589.001.0001

Csurka G (2017) A comprehensive survey on domain adaptation for visual applications. In: Domain Adaptation in Computer Vision Applications, Springer International Publishing, pp 1–35, DOI 10.1007/978-3-319-58347-1_1, URL https://doi.org/10.1007/978-3-319-58347-1_1

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, DOI 10.1109/cvpr.2009.5206848, URL https://doi.org/10.1109/cvpr.2009.5206848

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2014) Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning, pp 647–655

Duan L, Tsang IW, Xu D (2012) Domain transfer multiple kernel learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3):465–479, DOI 10.1109/tpami.2011.114, URL https://doi.org/10.1109/tpami.2011.114

van Eck NJ, Waltman L (2009) Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 84(2):523–538, DOI 10.1007/s11192-009-0146-3, URL https://doi.org/10.1007/s11192-009-0146-3

Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A (2009) The pascal visual object classes (VOC) challenge. International Journal of Computer Vision 88(2):303–338, DOI 10.1007/s11263-009-0275-4, URL https://doi.org/10.1007/s11263-009-0275-4

Fan J, Xu W, Wu Y, Gong Y (2010) Human tracking using convolutional neural networks. IEEE Transactions on Neural Networks 21(10):1610–1623, DOI 10.1109/tnn.2010.2066286, URL https://doi.org/10.1109/tnn.2010.2066286

Fernando B, Habrard A, Sebban M, Tuytelaars T (2013) Unsupervised visual domain adaptation using subspace alignment. In: 2013 IEEE International Conference on Computer Vision, IEEE, DOI 10.1109/iccv.2013.368, URL https://doi.org/10.1109/iccv.2013.368

Ge W, Yu Y (2017) Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, DOI 10.1109/cvpr.2017.9, URL

https://doi.org/10.1109/cvpr.2017.9

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, DOI 10.1109/cvpr.2014.81, URL https://doi.org/10.1109/cvpr.2014.81

Glorot X, Bordes A, Bengio Y (2011) Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 513–520

Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, DOI 10.1109/cvpr.2012.6247911, URL https://doi.org/10.1109/cvpr.2012.6247911

Guth F, de Campos TE (2018) Skin lesion segmentation using U-Net and good training strategies. Tech. rep., Cornell University Library, URL https://arxiv.org/abs/1903.06969, 1811.11314

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, DOI 10.1109/cvpr.2016.90, URL https://doi.org/10.1109/cvpr.2016.90

Hinton G, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Computation 18(7):1527–1554, DOI 10.1162/neco.2006.18.7.1527, URL https://doi.org/10.1162/neco.2006.18.7.1527

Hohman FM, Kahng M, Pienta R, Chau DH (2018) Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE transactions on visualization and computer graphics

Howard J, Ruder S (2018) Fine-tuned language models for text classification. CoRR abs/1801.06146, URL http://arxiv.org/abs/1801.06146, 1801.06146

III HD, Marcu D (2006) Domain adaptation for statistical classifiers. Journal of Artificial Intelligence Research 26:101–126, DOI 10.1613/jair.1872, URL https://doi.org/10.1613/jair.1872

Jia R, Liang P (2017) Adversarial examples for evaluating reading comprehension systems. CoRR abs/1707.07328, URL http://arxiv.org/abs/1707.07328, 1707.07328

Kessler JS (2017) Scattertext: a browser-based tool for visualizing how corpora differ. In: Proceedings of ACL-2017 System Demonstrations, Association for Computational Linguistics, Vancouver, Canada

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems 25, Curran Associates, Inc., pp 1097–1105, URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

L Fei-Fei JD (2017) Where have we been? Where are we going? URL http://image-net.org/challenges/talks_2017/imagenet_ilsvrc2017_v1.0.pdf, [Online; a june/28 2018]

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444, DOI 10.1038/nature14539, URL https://doi.org/10.1038/nature14539

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, DOI 10.1109/cvpr.2015.7298965, URL https://doi.org/10.1109/cvpr.2015.7298965

Lucena O, Junior A, Moia V, Souza R, Valle E, Lotufo R (2017) Transfer learning using convolutional neural networks for face anti-spoofing. In: Lecture Notes in Computer Sci-

ence, Springer International Publishing, pp 27–34, DOI 10.1007/978-3-319-59876-5_4, URL https://doi.org/10.1007/978-3-319-59876-5_4

Mahajan D, Girshick R, Ramanathan V, He K, Paluri M, Li Y, Bharambe A, van der Maaten L (2018) Exploring the limits of weakly supervised pretraining. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 181–196

Mariano AM, Rocha MS (2017) Revisão da literatura: Apresentação de uma abordagem integradora. In: XXVI Congreso Internacional de la Academia Europea de Dirección y Economía de la Empresa (AEDEM), Reggio Calabria, vol 26

Ng A (2016) Nuts and bolts of applying deep learning. URL https://www.youtube.com/watch?v=F1ka6a13S9I, accessed online on Nov/20/2018

Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA, CVPR, pp 1717–1724, DOI 10.1109/CVPR.2014.222, URL https://doi.org/10.1109/CVPR.2014.222

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans on Knowl and Data Eng 22(10):1345–1359, DOI 10.1109/TKDE.2009.191, URL http://dx.doi.org/10.1109/TKDE.2009.191

Pan SJ, Tsang IW, Kwok JT, Yang Q (2011) Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks 22(2):199–210, DOI 10.1109/tnn.2010.2091281, URL https://doi.org/10.1109/tnn.2010.2091281

Park I, Yoon B (2018) Identifying promising research frontiers of pattern recognition through bibliometric analysis. Sustainability 10(11):4055

Raina R, Battle A, Lee H, Packer B, Ng AY (2007) Self-taught learning. In: Proceedings of the 24th international conference on Machine learning - ICML '07, ACM Press, DOI 10.1145/1273496.1273592, URL https://doi.org/10.1145/1273496.1273592

Rezende ERD, Ruppert GC, Carvalho T (2017) Detecting computer generated images with deep convolutional neural networks. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, DOI 10.1109/sibgrapi.2017.16, URL https://doi.org/10.1109/sibgrapi.2017.16

Ruder S (2019) Neural transfer learning for natural language processing. PhD thesis, National University of Ireland, Galway

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115(3):211–252, DOI 10.1007/s11263-015-0816-y, URL https://doi.org/10.1007/s11263-015-0816-y

Si S, Tao D, Geng B (2010) Bregman divergence-based regularization for transfer subspace learning. IEEE Transactions on Knowledge and Data Engineering 22(7):929–942, DOI 10.1109/tkde.2009.126, URL https://doi.org/10.1109/tkde.2009.126

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Taylor ME, Stone P (2009) Transfer learning for reinforcement learning domains: A survey. J Mach Learn Res 10:1633–1685, URL http://dl.acm.org/citation.cfm?id=1577069.1755839

Thrun S (1996) Is learning the n-th thing any easier than learning the first? In: Advances in neural information processing systems, pp 640–646

Torrey L, Shavlik J (2010) Transfer learning. In: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, IGI Global, pp 242–264

Vapnik VN (1998) Statistical Learning Theory. Wiley-Interscience

Venkateswara H, Eusebio J, Chakraborty S, Panchanathan S (2017) Deep hashing network for unsupervised domain adaptation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, DOI 10.1109/cvpr.2017.572, URL https://doi.org/10.1109/cvpr.2017.572

Vogel R, Güttel WH (2012) The dynamic capability view in strategic management: A bibliometric review. International Journal of Management Reviews pp n/a–n/a, DOI 10.1111/ijmr.12000, URL https://doi.org/10.1111/ijmr.12000

Yang J, Yan R, Hauptmann AG (2007) Cross-domain video concept detection using adaptive svms. In: Proceedings of the 15th international conference on Multimedia, ACM Press, DOI 10.1145/1291233.1291276, URL https://doi.org/10.1145/1291233.1291276

Zamir AR, Sax A, Shen WB, Guibas LJ, Malik J, Savarese S (2018) Taskonomy: Disentangling task transfer learning. CoRR abs/1804.08328, URL http://arxiv.org/abs/1804.08328, 1804.08328

Zhao M, Li T, Abu Alsheikh M, Tian Y, Zhao H, Torralba A, Katabi D (2018) Through-wall human pose estimation using radio signals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7356–7365

Zhu J, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR abs/1703.10593, URL http://arxiv.org/abs/1703.10593, 1703.10593

B AN INFORMATION THEORETICAL TRANSFERABILITY METRIC

[version: June 19, 2020 at 18:52]

An Information Theoretical Transferability Metric

Fred Guth
fredguth@fredguth.com
Departamento de Ciência de Computação, Universidade de Brasília
Brasília, DF, Brazil

ABSTRACT

Deep Learning in Computer Vision deserves a lot of its success to the simple Transfer Learning technique of fine-tunning, using a pre-trained model as feature extractor for a new task, which unlocks the floodgates of very big datasets like ImageNet[12] even to the most modest classification task. It is so ubiquitous that not doing it is considered foolhardy [15]. Despite of that, little is known on when and to what extend fine-tunning works. Decisions on which pre-trained model to use are *ad hoc* and automatic model selection seems far from reality. Fortunately, a recent theoretical effort [1, 2, 4–9, 14, 17–20, 25] have shed some light upon these matters using Information Theory. An interesting result is that Fisher Information Matrices can be used as task embeddings and distances between them can be seen as the properties of "closeness" and "transferability". In this work, we present an (hopefully) accessible review of this theoretical *tour de force*.

CCS CONCEPTS

• **Computing methodologies** → **Transfer learning**.

KEYWORDS

transfer learning, automation, deep learning, complexity, task2vec, information theory, learning theory

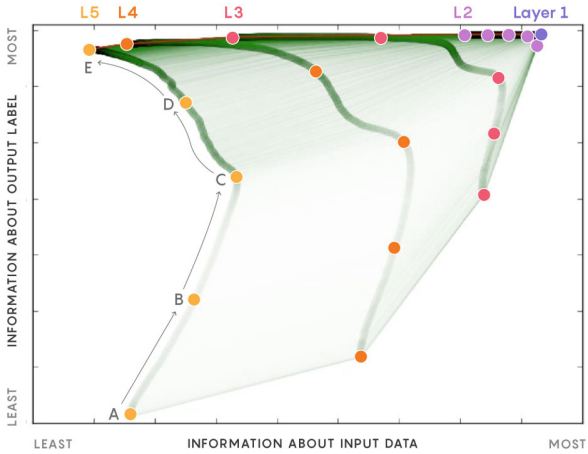
1 INTRODUCTION

Despite all success of supervised Deep Learning in a myriad of applications, humans still show a much better generalization ability from very few samples (Goodfellow et al. [11]). The huge demand for well labeled data, which are expensive and difficult to obtain, lead to a growing interest for Transfer Learning, which harness previously acquired knowledge to new tasks so that they can be learned in a more effective and efficient way. It is notoriously known that one of the main reasons for deep learning success is due to the fact that transfer learning works very well for it.

Indeed, Transfer learning became so ubiquitous in the Deep Learning context that Mahajan et al. [15] claims that training from *tabula rasa*, instead of using a pre-trained ImageNet[12] model as feature extractor, is considered foolhardy in Computer Vision.

A recent survey on Research Frontiers in Transfer Learning [13] points that this simple transfer learning technique, also known as fine-tunning, is the most used approach in recent publications. This confirms Torrey and Shavlik [21] view that transfer has been applied on an *ad hoc* basis. Moreover, this context imply a new constraint on deep learning usage in practice: experts capable of selecting models and datasets to build feature extractors for new tasks.

This same survey presents as open problems the lack of:



A **INITIAL STATE:** Neurons in Layer 1 encode everything about the input data, including all information about its label. Neurons in the highest layers are in a nearly random state bearing little to no relationship to the data or its label.

B **FITTING PHASE:** As deep learning begins, neurons in higher layers gain information about the input and get better at fitting labels to it.

C **PHASE CHANGE:** The layers suddenly shift gears and start to "forget" information about the input.

D **COMPRESSION PHASE:** Higher layers compress their representation of the input data, keeping what is most relevant to the output label. They get better at predicting the label.

E **FINAL STATE:** The last layer achieves an optimal balance of accuracy and compression, retaining only what is needed to predict the label.

Source: Reproduced from Wolchover [23].

Figure 1: Illustration of Tishby's conjecture in the Information Plane.

(a) theory to support why, when and to what extent fine-tunning works;
(b) transferability metrics, which could lead to meta-learning algorithms that automatic select the most promising pre-trained model for a new task;

This two open problems are the subject of this paper.

The best paper at CVPR 2018 (Zamir et al. [24]) proposes a experimental evaluation of transferability, creating a taxonomic map for task transfer learning. This leaves open the possibility of creating

meta-learning algorithms to automatically select models for new tasks based on previous empirical results.

The method is highly computational intensive, though. The total number of transfer functions trained to build a taxonomy of 26 vision tasks was around 3000 and took 47,886 GPU hours[24, §4 Experiments].

It is desired for such metric to be more easily computed. One recent approach that promises to fill this gap is Achille et al. [2, Task2Vec], where the authors use the Fisher Information Matrix (FIM), a method widely used in physics to forecast empirical results of hypothesis, to generate a task embedding, which conveys the task complexity and can be used to generate transferability metrics. Unfortunately, the AWS* sponsored research did not make available its implementation and experimental data as open-source.

1.1 Overview

In sections 2 and 3, we give some background and motivation for sections 4 and 5, which present the theory behind task embeddings and are the crux of the paper. Section 6 presents the application of the theory that motivated our interest in the first place, how to use information theoretical supported transferability measures. Section 7 summarize the findings and suggest future venues of inquiry.

1.2 Related Work

A superficial and joyful overview of the Information Bottleneck Theory can be found in Wolchover [23]. For a more in depth and broad overview of IB and its applications we refer to Hafez-Kolahi and Kasaei [14]. Cover and Thomas [10] is an superb source on the foundations of Information Theory and covers most (if not all) concepts used here in a very pleasant way. For more on information/task complexity we suggest [5, 8].

2 PRELIMINARIES AND NOTATIONS

Let $X \in \mathcal{X}$ be a random variable of the input space (e.g., an image) and $Y \in \mathcal{Y}$ a random variable that we want to infer (e.g, a label), which is therefore referred as our *task*. We consider that X and Y are continuous, but represented in a finite precision machine, therefore quantized into discrete values. A dataset is a finite set of m samples $\mathcal{D} = \{(x_i, y_i)\}^m$.

We will make frequent use of the following basic information theoretic concepts[10]: $H(X) = \mathbb{E}_p[\log \frac{1}{p(x)}]$ † is the Shannon entropy of the random variable X, not to be confused with $\mathcal{H}$, the hypothesis space in classical learning theory. The conditional entropy is $H(X|Y) = \mathbb{E}_{\hat{y}} H(X|Y = \hat{y}) = H(X, Y) - H(Y)$. We denote $p(X, Y)$ as the joint probability of X and Y, and the corresponding mutual information (MI or information gain) $I(X; Y) = I(Y; X)$ is defined as:

$$I(X;Y) = \mathrm{KL}(\, p(Y,X) \,\|\, p(Y)p(X) \,) = \mathbb{E}_X \,\mathrm{KL}(\, p(Y|X) \,\|\, p(Y) \,), \quad (1)$$

where $\mathrm{KL}(\, p \,\|\, q \,) = \mathbb{E}_p[log\frac{p}{q}]$ is the Kullbach-Leibler divergence (KL-divergence) between the the probability distributions p and q, which normally denote the true distribution of a variable and the estimated distribution respectively. Cross-entropy is:

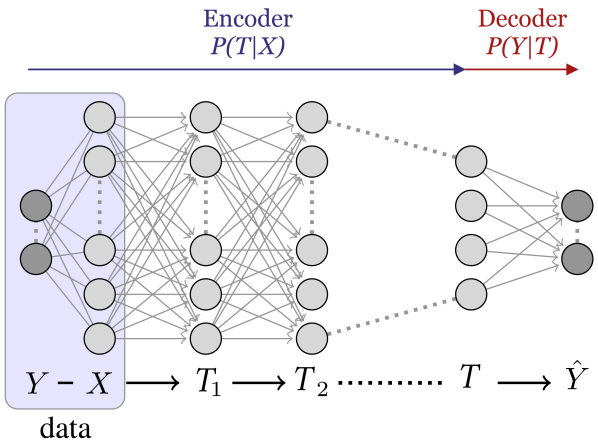
$$H_{p,q}(X) = \mathbb{E}_p[-\log q(X)] = H_p(X) + \mathrm{KL}(\, p(X) \,\|\, q(X) \,). \quad (2)$$

Total correlation TC(T), a.k.a. the multi-variate mutual information is:

$$TC(T) = \mathrm{KL}(\, p(T) \,\|\, \textstyle\prod_i p(T_i) \,)$$

$$TC = 0 \iff T_1 \perp T_2 \perp \ldots \perp T_k,$$

where $p(T_i)$ are the marginal distributions of the components of T. A DNN model (figure 2) is a succession of K layers, each k^{th} layer denoted as $T_k = \phi_k(W_k T_{k-1})$, where W_k represent the weight vector and ϕ_k a non-linear function (usually ReLU) in the k^{th} layer.



Source: Adapted from Shwartz-Ziv and Tishby [17].

Figure 2: Deep Neural Network layers as representations of the input.

Given the weights, the layers form a Markov chain of successive internal representations of the input variable $X : Y \to X \to T_1 \to T_i \to \ldots \to T_K \to \hat{Y}$, i.e. $p(T|X, Y) = p(T|X)$ and their MI obey a chain of Data Processing Inequalities (DPI)[10, 17], therefore $I(Y; X) \geq I(Y; T)$. The DNN can be seen as an encoder from $X \to T$, $p(T|X)$, and a decoder from $T \to \hat{Y}$, $p(Y|T)$.

3 LEARNING THEORY HAS FAILED DEEP

Learning Theory studies the complexity of learning algorithms, not only in terms of steps but also in the number of samples needed to guarantee true error bounds (generalization bounds). It started with the seminal work of Valiant [22] introducing the PAC Learning model.

The generalization bounds obtained by PAC Learning is [16, Theorem 2.13]:

$$\varepsilon^2 < \frac{\log |\mathcal{H}_\epsilon| + \log \delta^{-1}}{2m} \quad (3)$$

where:

ε is the tolerance margin between training and generalization errors. The generalization error measures the accuracy of the algorithm for previously unseen data, the test error;

* Amazon Web Services † Despite the notation, it is not a function of X, but a function its distribution p(X)

$|\mathcal{H}_\epsilon|$ is the cardinality of the ϵ-cover of the hypothesis space. Typically, it is assumed that $|\mathcal{H}_\epsilon| \sim \frac{1}{\epsilon}^d$;

δ is the confidence margin;

m is the number of training samples, a.k.a. the sample complexity;

d is the Vapnik–Chervonenkis dimension of the hypothesis space. In the case of neural networks it is $O(|\theta| log |\theta|)$, where $|\theta|$ is the number of parameters in the network.

The generalization error is bounded by a function of the hypothesis space and the dataset sizes.

3.1 Criticism

The main criticism on the current state of Learning Theory in the context of DNNs are:

(1) the bounds are too loose and, therefore, not very valuable in practice;
(2) it depends on the model (size of hypothesis space), not only on the problem;
(3) its preference for simpler algorithms (smaller hypothesis space) does not explain the fact that larger and deeper Deep Neural Networks (DNNs) usually achieve better accuracy and generalization.

4 AN INFORMATION THEORY OF DEEP LEARNING

Tishby and Zaslavsky [20] propose a new Learning Theory of Deep Learning based on Shannon's Information Theory. Therefore, it is important to establish some context.

Tasks can be easy or difficult depending on how information is represented, a classical example is that it is much easier to calculate using hindu-arabic numerals than with roman numerals. Thus, it is reasonable to think of supervised trained DNNs as performing representation learning, where the last layer, the head, is typically a softmax regression classifier and all previous layers just learn to provide a good representation for this last classifier[11, Chapter 15].

Another way is to think of the last layer as decoding a message that was encoded by the rest of the network (see figure 2). In this view, each layer can be seen as a single random variable, T_i, and the network as the communication channel itself, $p(T|X)$, to which all Shannon information properties apply.

4.1 Desiderata for representations

Let T denote a representation of X, that is optimal to the task Y, meaning that T captures and exposes only the information from X which is relevant to Y. Ideally, this representation should be [6]:

a statistic: a function $T \sim p(T|X)$;

sufficient: $I(Y;T) = I(Y;X)$, so there is no loss in relevant Y information;

minimal: $I(T;X)$ is minimized, so that it retains as little of X as possible. This means there is an encoding from X to T that keep only relevant information;

invariant: to the effect of nuisances N, where $N \perp Y \rightarrow I(N;Y) = 0 \rightarrow I(T;N) = 0$ means that if N does not have information about Y, there should not be information of N in the representation T, otherwise the classifier could fit to spurious correlations;

maximally disentangled: no information will be present in the correlations between components of T.

4.2 The Information Bottleneck

The Information Bottleneck is a method for finding minimal sufficient statistics developed by Tishby et al. [19]:

$$\begin{aligned} T^* = \arg\min_T \quad & I(T;X) \\ \text{s.t.} \quad & I(T;Y) = I(X;Y) \end{aligned} \tag{4}$$

Applying the lagrangian relaxation, we have:

$$\begin{aligned} T^* &= \arg\min_T \mathcal{L} \\ \mathcal{L} &= \min_{q(T|X)} I(T;X) - \beta\, I(T;Y), \beta > 0 \end{aligned} \tag{IB}$$

where β is the Lagrange multiplier. Tishby and Zaslavsky [20] used the Information Bottleneck (IB) to formulate the deep learning goal as an information trade-off between sufficiency and minimality, accuracy and generalization, prediction and compression.

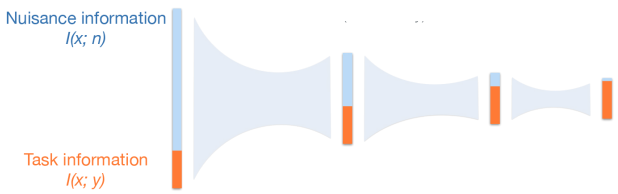
4.3 Emerging Properties of DNNs

It is interesting to notice that it is possible to rewrite (Appendix A.1) the IB formulation as:

$$\mathcal{L} = \min_q \underbrace{H_{p,q}}_{\text{cross-entropy}} (Y|T) + \beta \underbrace{I(T;X)}_{\text{regularizer}}, \beta > 0 \quad \text{(IB Lagrangian)}$$

and the cross-entropy, the most successful loss function for classification tasks, naturally emerges, as does a not usual regularizer in a second term.

4.3.1 Invariance and minimality. Achille and Soatto [6] demonstrate that by using SGD there is in fact an implicit compression of information, showing the regularizer is there, but implicit. Also, they show that by enforcing the minimization of the information about the input representation $I(T;X)$, invariance and disentanglement naturally emerges as well, satisfying the desiderata(§4.1).



Source: Adapted from Achille [1].

Figure 3: Stacking layers improve generalization.

The intuition for the emergence of invariance is quite simple to understand. The architecture enforces a dimensionality reduction by stacking layers and/or applying pooling, while the algorithm is trying to maintain the information about the labels (figure 3). Besides, techniques like dropout and batch-normalization add noise to the training. Noise in the communication channel reduces its capacity, therefore, reduces the upper bound to $I(X;T)$. If $I(X;T)$ is being reduced while $I(Y;T)$ is maintained, what is being reduced is $I(T;N)$, where $N \perp Y$.

Another way to reduce the information is to train with the regularizer of eq. (IB Lagrangian) explicitly. This is what Achille and Soatto [7] propose.

4.4 Tishby's Conjecture: Learning is forgetting

Shwartz-Ziv and Tishby [17] conjectures that DNNs work in two phases: (1) a **Fitting Phase** where the DNN rapidly fit to the training labels; and (2) a **Compression Phase** where it spends most of the time, compressing the inputs and therefore "forgetting" as much of X that it can, without losing information about Y.

To support this view, they present the Information Plane (Figure 1), where darker colors represent later epochs in the training (and the darkening of the colors is in constant rate to the number of epochs). Its experimental setup, however, has been challenged and no consensus has been reached [8]. Regardless, Tishby's conjecture is the basis of a new Learning Theory that is worth attention.

4.5 Information Theoretical Bounds

As the current Learning Theory does not explain well documented observed facts §3.1, Tishby and Zaslavsky [20] propose a change of focus from worse case model-dependent distribution-independent PAC bounds to typical data-dependent model-free bounds. In other words, instead of bounding using the expressivity of the hypothesis space ($\mathcal{H}$), bounding by the limits of the compression of the input data.

In this new information learning theory, the complexity of a learning problem only depends on the problem itself, the data and the information it conveys, and not on any property of the algorithm chosen to solve it, which is a very interesting proposition.

In the IB formulation, $p(T|X)$ is a stochastic mapping between each value $x \in \mathcal{X}$ to a value $t \in \mathcal{T}$, inducing a partition of $\mathcal{X}$, a quantization of $\mathcal{X}$ into a codebook $\mathcal{T}$. In terms of Shannon's Information Theory, it can be seen as a communication from X to T.

A reliable communication is one that require that different input sequences produce disjoint output sequences. As $|\mathcal{X}|$ is quite large, we can use the concept of typicality. The typical set, $\mathcal{T}_\epsilon$, is the set of distributions that are near the true distribution. Empirical probability distributions that are non typical have exponentially smaller probability. The current learning theory works for any distribution of data, but $|\mathcal{T}_\epsilon| \ll |\mathcal{X}|$.

A typical n-sequence of T symbols can, from the ϵ-partition property (AEP, [10, Theorem 3.1.2]), represent $\sim 2^{n}H(X|T)$ possible X n-sequences. And again from AEP, there are $\sim 2^{n}H(X)$ different n-sequences of X. Therefore, as we have to ensure that no two sequences of X maps to the same sequence of T, the total number of disjoint subsets T_i of X is upper bounded by $2^{n(H(X)-H(X|T)} = 2^{nI(X;T)}$ [18]. From equation 3, we now can bound the generalization error. Let T_ϵ be the space of ϵ-partitions of X:

$$
\begin{aligned}
|\mathcal{H}| \leq 2^{|\mathcal{X}|} &\rightarrow 2^{|\mathcal{T}_\epsilon|} \\
|\mathcal{T}_\epsilon| &= 2^{I(T;X)} \\
\epsilon^2 &< \frac{2^{I(T;X)} + log\delta^{-1}}{2m}
\end{aligned} \tag{5}
$$

Every k bits of compression have the same effect to the error as 2^k samples.

4.6 IB Achille's heel

The formulation presented in eq. (IB Lagrangian) is a representation of yet not observed future data (the activations of future X). It is sort of a wishful thinking, as we only have past data.

A common criticism to IB is that a totally valid minimization of $I(T;X)$ is to memorize the index of input training data that map to a certain label. Although keeping little information, this obviously will not generalize. We need a function obtained in training that is guaranteed to work in test time.

This is basically what Achille and Soatto [6] propose with the Information in the Weights Bottleneck.

5 INFORMATION IN THE WEIGHTS

5.1 Deep Learning Reality

Deep Learning is usually associated with DNNs, but it is only one of its components:

(1) DNN architecture
(2) Optimizer (SGD)
(3) Dataset
(4) Loss function, usually:

$$
\begin{aligned}
\mathcal{L}(W) &= H_{p,q}(Y|X,W) \\
&= H_{p,q}(D|W)
\end{aligned}
$$

where, as a reminder, p is the true distribution and q is the model's approximation of the true distribution.

Not only the architecture is important to current Deep Learning success. As it was already mentioned and will be seen shortly, in this information theoretical view the stochastic of SGD plays a crucial role, so does the choice of cross-entropy for the loss function and the use of large datasets.

A known problem, though, is that DNNs are prone to overfitting. Actually, Zhang et al. [26] shows that state-of-the-art convolutional deep neural networks can easily fit a random labeling of training data.

5.2 Overfitting

To understand why DNNs overfit, Achille and Soatto [6] propose to decompose the cross-entropy. Assuming the dataset is sampled from some generative model $p(D|\theta)$:

$$
\boldsymbol{H_{p,q}(D|W)} = H_p(D,\theta) + I(\theta;D|W) + \mathbb{E}\,\mathrm{KL}(\,p\,\|\,q\,) - \underbrace{I(D;W|\theta)}_{overfitting}
$$

A naive idea to eliminate overfitting would be to rewrite the loss to eliminate the overfitting term:

$$\mathcal{L}(W) = H_{p,q}(D|W) + I(D;W|\theta)$$

To calculate $I(D;W|\theta)$, true distribution p_θ is needed, which we are just trying to approximate with q in training. Hence we are presented with chicken-egg problem. Rather, one can add a Lagrangian multiplier to upper bound $I(D;W|\theta)$:

$$\boldsymbol{\mathcal{L}(W) = H_{p,q}(D|W) + \underbrace{\beta I(D;W)}_{\text{Information in the Weights}}} \qquad \text{(new IBL)}$$

Remarkably, this has the same form of the IB Lagrangian equation.

5.3 A new Information Lagrangian

The question now is if this new Lagrangian emerged from trying to eliminate overfitting is somehow related to the IB Lagrangian in the activations presented before.

$$\underset{\text{input}}{X} \longrightarrow \underset{\text{activations}}{T} \longrightarrow \underset{\text{label}}{Y}$$

$$\min_{q(T|X)} \mathcal{L}(W) = H_{p,q}(Y|T) + I(T;X) \qquad \text{(Activations IB)}$$

$$\underset{\text{dataset}}{D} \longrightarrow \underset{\text{weights}}{W} \longrightarrow \underset{\text{real distribution}}{p_{(Y|X)}}$$

$$\min_{q(T|X)} \mathcal{L}(W) = H_{p,q}(D|W) + I(D;W) \qquad \text{(Weights IB)}$$

Intuitively, the *Weights IB* seems to be the dual to the *Activations IB*, as $I(T;X)$ which measures the complexity of the activations representation, can be defined by the amount of weight in the network: low or zero weights will connect to the activations that are not in T^* which minimizes $I(T;X)$.

Achille and Soatto [6, Corollary C.8] have proved that indeed $I(T;X) \leq I(W;D)$. In other words, the amount of information needed to be memorized to minimally represent the dataset is an upper bound to the amount of information needed to guarantee invariance (a small error in test data).

As $I(W;D)$ can be calculated, this development allows one to explicitly regularize the training. That is exactly what [7, Information Dropout] proposes.

5.4 Shannon vs. Fisher Information

By using eq. (1), eq. (new IBL) can be rewritten as:

$$\boldsymbol{\mathcal{L}(W) = H_{p,q}(D|W) + \beta\,\mathrm{KL}\Big(\underbrace{q(W|D)}_{\text{training output}} \;\|\; \underbrace{p(W)}_{\text{fixed prior}}\Big)}$$

In other words, $I(W;D)$ is the divergence of the conditional model distribution $q(W|D)$ and the expected prior averaging all datasets. If, instead, we assume an isotropic gaussian[‡] as the prior, the information in the weights when W_* is minimal, is given by:

$$\mathrm{KL}(\, q(W|D) \,\|\, p(W)\,) = \frac{1}{2}\left(\log|\mathcal{F}m| + \cancel{\log \lambda^2 I} + \frac{\cancel{W_*^2}}{\cancel{\lambda^2 I}}\right),$$

where the canceled terms are the ones that do not depend on $q(W|D)$ and can be ignored, $\log|\mathcal{F}|$ is the log-determinant of Fisher Information Matrix of the weights and m is the number of samples in the dataset.

This is quite interesting as it gives us an analytical and fast calculation of a bound to $I(W;D)$:

$$I(T;X) \leq I(D;W) \leq \log|\mathcal{F}(W^*)| \qquad (6)$$

5.5 Information, Flat Minima and Generalization

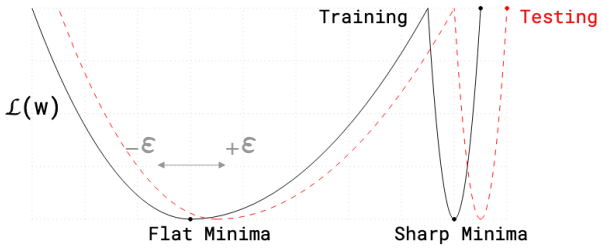


Figure 4: Information in the weights explain the preference of SGD for flat minima.

The relation of low Fisher Information in the weights and good generalization in the activations bring us another interesting insight. It is a well documented phenomena that DNNs trained with SGD tend to find flat minima, meaning saddle points where some noise in the test data will minimally affect the loss (see figure 4). Here we have a theoretical reason why this is happening. SGD is implicitly regularizing the information in the weights and that is the same as looking for minimas with low Fisher Information, which are Flat Minimas.

6 TOPOLOGY OF LEARNING TASKS

To this day, transferability is either measured experimentally [24] or inferred subjectively by experts according to tasks "proximity". Given an analytical transferability measure, obtained directly from the data in a cost-effective way, with experimentally proved prediction ability, automatic selection of source tasks as feature extractors for target tasks (auto-DL) is a simple search in the topology of learning tasks.

This illustrates the importance of building such topology. In other words, we need to know:

- What is the complexity of a learning task?
- How far or close are two tasks?
- How difficult it is to transfer from one task to another?

The aim of this section is to show such a transferability metric based on the information theory developments explained in previous sections.

6.1 Complexity of Learning Tasks

Intuitively, the complexity of a learning task is related to its best expected true error. This is exactly what what the Information in the Weights (section 5) give us.

[‡] An isotropic gaussian is one where the covariance matrix is represented by $\Sigma = \lambda^2 I$.

From eq. (5) and eq. (6), we have:

$$\epsilon^2 \propto \frac{2^{I(X;T)}}{2m} \propto \frac{2^{\log|\mathcal{F}|)}}{2m} \propto \frac{|\mathcal{F}|}{m}$$

Given a fixed architecture, the amount of information in the weight (Fisher Information) measures how much "memorization" was used to fit the model. High information in the weights suggest more "difficult" tasks. The Fisher Matrix (FIM) measure the resilience of the loss due to perturbation in the weights (see figure 4). If a weight accept more noise, it is less important and there is no need to "memorize" it. Also, as seen in 5.5, this amount of noise has direct correspondence to generalization. Using this intuition, [3, Task2Vec] uses the diagonal of the FIM as an embedding that represent the task itself. Since the FIM can be too noisy when trained from few examples, the diagonal of the FIM is used as it is considered a more simple and robust representation.

Different choices of fixed architectures, however, produce FIMs that are not comparable. To address this, a standard "probe" network pre-trained on ImageNet is used. The FIM of the probe represents the canonical task t_0 from which other tasks are compared. The embedding of a new task t_i is obtained by re-training only the classifier layer $p(T;Y)$, which usually can be done efficiently, and then computing the FIM for the feature extractor parameters.

6.2 Measures in the Task Topology

Achille et al. [3, Task2Vec] propose two measures in the task topology: (1) Relatedness, a semantic distance which they call taxonomic distance; and (2) Transferability, which they call transfer distance.

6.2.1 Relatedness: semantic distance. Experts base their subjective *ad hoc* choices of which tasks to use as source for a specific target to the idea of "proximity" or "closeness" of tasks (how far or close are two tasks). Thus, a measure of semantic distance with clear operational meaning is highly desired, that metric is the symmetric distance of task embeddings:

Definition 6.1 (Symmetric distance). *The symmetric distance between normalized embeddings measures its "closeness":*

$$d_{sym}(F_a, F_b) = d_{cos}\Big(\frac{F_a}{F_a+F_b}, \frac{F_b}{F_a+F_b}\Big),$$

where d_{cos} is the cosine distance, F_a and F_b are the two task embeddings, and the division is element-wise.

6.2.2 Transferability. Transferability (or fine-tunning gain) from a task t_a to a task t_b is the difference in expected performance between a model trained for task b from a fixed initialization, t_0, and the performance of a solution to t_a fine-tuned for t_b:

$$D_{\text{f-t}}(t_a \to t_b) = \frac{\mathbb{E}[\ell_{a\to b}] - \mathbb{E}[\ell_b]}{\mathbb{E}[\ell_b]},$$

where expectations are taken over all trainings, ℓ_b is the final test error obtained by training task b from initialization, and $\ell_{a\to b}$ is the error when starting from a solution to task a fine-tuned for task b. Hence, transferability depends on the similarity between two tasks and the complexity of the first. Indeed, the fact that pre-training in ImageNet has become a *de facto* standard [15] is due to its high complexity.

This intuition suggest the following measure for transferability:

Definition 6.2 (Asymmetric distance). *The asymmetric distance*[§] *between two normalized embeddings measures its transfer gain of using one as the source of the other:*

$$d_{asym}(t_a \to t_b) = d_{sym}(t_a, t_b) - \alpha d_{sym}(t_a, t_0),$$

where t_0 is the canonical embedding, and α is an hyperparameter.

In their experiments, the best value of α ($\alpha = 0.15$ when using a ResNet-34 pre-trained on ImageNet as the probe network) was considered robust to the feature-extractor selection meta-task.

7 DISCUSSION

In this work we presented the information theoretical support for the use of Fisher Information Matrices as task embeddings with which it is possible to measure the closeness and the transferability between tasks.

This theory is full of counter-intuitive and elucidative explanations for several puzzling phenomena. It shows that:

a) Noise in training and architectural bottlenecks are helpful;
b) Stacking layers increase generalization;
c) There is no "curse of dimensionality", as the complexity of a task is related not to the number of parameters of a model, but only to the compression limits of the data itself;
d) Reducing information in the weights during training yields smaller error in the activations in testing;
e) Tishby's conjecture is right and learning is forgetting, what maybe his information plane cannot prove, but certainly Achille and Soatto proof of the bound of the information in the weight to the information in the activation does;
f) The preference of SGD for flat minima can be explained by an implicit regularizer that tries to constraint the amount of information in the weights;
g) A FIM can be used as a representation of a Task;
h) It is possible to calculate asymmetric "distances" between tasks as transferability metrics in a cost efficient way;
i) There is a potential path towards deep learning automation.

At the same time the this theoretical *tour de force* lifts the veil of several aspects of deep learning, it also opens new venues of inquiry like how to use the Fisher Information Matrix to not only decide task sources but also which layer has the right amount of semantic for the target task in question; how does the choice of the learning rate affect the noise in the channel and if it can explain the phenomenon of "super-convergence"; How good is the analytical measures of complexity and measurability proposed compared to the empirically obtained ones in Zamir et al. [24]. These are questions that will certainly be subject of our future investigations.

REFERENCES

[1] Alessandro Achille. 2019. CS103 - Topics in Representation Learning, Information Theory and Control. https://alexachi.github.io/cs103/index.html [Online; accessed on June 20th, 2019].
[2] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2Vec: Task Embedding for Meta-Learning. *arXiv preprint arXiv:1902.03545* (2019).
[3] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. 2019.

[§] Not properly a distance as asymmetric distance is an oxymoron and it also can be negative.

Task2Vec: Task Embedding for Meta-Learning. *CoRR* abs/1902.03545 (2019). arXiv:1902.03545 http://arxiv.org/abs/1902.03545
[4] Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. 2019. The Information Complexity of Learning Tasks, their Structure and their Distance. *arXiv preprint arXiv:1904.03292* (2019).
[5] Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. 2019. The Information Complexity of Learning Tasks, their Structure and their Distance. *CoRR* abs/1904.03292 (2019). arXiv:1904.03292 http://arxiv.org/abs/1904.03292
[6] Alessandro Achille and Stefano Soatto. 2017. On the Emergence of Invariance and Disentangling in Deep Representations. *CoRR* abs/1706.01350 (2017). arXiv:1706.01350 http://arxiv.org/abs/1706.01350
[7] Alessandro Achille and Stefano Soatto. 2018. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence* 40, 12 (2018), 2897–2905.
[8] Alessandro Achille and Stefano Soatto. 2019. Where is the Information in a Deep Neural Network? arXiv:cs.LG/1905.12213
[9] William Bialek, Ilya Nemenman, and Naftali Tishby. 2001. Predictability, complexity, and learning. *Neural computation* 13, 11 (2001), 2409–2463.
[10] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA.
[11] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org/
[12] M. Guillaumin and V. Ferrari. 2012. Large-scale knowledge transfer for object localization in ImageNet. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/cvpr.2012.6248055
[13] Fred Guth. 2019. *Research Frontiers in Transfer Learning: a systematic review with a quantitative approach*. Technical Report. UnB.
[14] Hassan Hafez-Kolahi and Shohreh Kasaei. 2019. Information Bottleneck and its Applications in Deep Learning. *arXiv preprint arXiv:1904.03743* (2019).
[15] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 181–196.
[16] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. The MIT Press.
[17] Ravid Shwartz-Ziv and Naftali Tishby. 2017. Opening the Black Box of Deep Neural Networks via Information. *CoRR* abs/1703.00810 (2017). arXiv:1703.00810 http://arxiv.org/abs/1703.00810
[18] Noam Slonim. 2002. *The information bottleneck: Theory and applications*. Ph.D. Dissertation. Hebrew University.
[19] Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The Information Bottleneck Method. 368–377.
[20] Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*. IEEE, 1–5.
[21] Lisa Torrey and Jude Shavlik. 2010. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 242–264.
[22] L. G. Valiant. 1984. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing - 84*. ACM Press. https://doi.org/10.1145/800057.808710
[23] Natalie Wolchover. 2017. New Theory Cracks Open the Black Box of Deep Learning. https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-learning-20170921/
[24] Amir Roshan Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling Task Transfer Learning. *CoRR* abs/1804.08328 (2018). arXiv:1804.08328 http://arxiv.org/abs/1804.08328
[25] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences* 115, 31 (2018), 7937–7942.
[26] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *CoRR* abs/1611.03530 (2016). arXiv:1611.03530 http://arxiv.org/abs/1611.03530

Appendices

A PROOFS

A.1 Information Bottleneck Lagrangian

$$\begin{aligned} \min_{p(T|X)} \quad & I(T;X) \\ \text{s.t.} \quad & I(T;Y) \le I(X;Y) \end{aligned}$$

Can be rewritten:

$$\begin{aligned} I(T;Y) &\le I(X;Y) \\ H(Y) - H(Y|T) &\le H(Y) - H(Y|X) \\ H(Y|X) - H(Y|T) &\le 0 \end{aligned}$$

$$\begin{aligned} \min_{p(T|X)} \quad & I(T;X) \\ \text{s.t.} \quad & H(Y|X) - H(Y|T) \le 0 \end{aligned}$$

Using the Lagrangian multiplier, it can be relaxed as:

$$\gamma > 0, \quad \min_{p(T|X)} \quad I(T;X) + \gamma H(Y|T)$$

$$\frac{1}{\gamma} = \beta > 0, \quad \min_{p(T|X)} \quad H(Y|T) + \beta I(T;X) \qquad \square$$

A.2 Cross-Entropy Decomposition

We want to prove that:

$$H_{p,q}(D|W) = H_p(D,\theta) + I(\theta;D|W) + \mathbb{E}\,\mathrm{KL}(\,p \,\|\, q\,) - I(D;W|\theta)$$

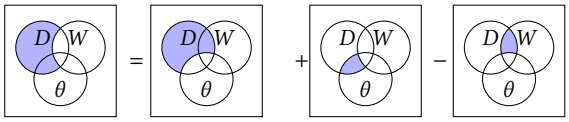
From eq. eq. (2):

$$H_{p,q}(D|W) = H_p(D|W) + \mathrm{KL}(\,p(D|W) \,\|\, q(D|W)\,)$$

Therefore, we only need to prove that:

$$H_p(D|W) = H_p(D|\theta) + I(D|W;\theta) - I(D;W|\theta)$$

Which is clear with the help of the following Venn diagrams:



$$H_p(D|W) = H_p(D|\theta) \qquad +I(D|W;\theta) - I(D;W|\theta) \qquad \square$$

BIBLIOGRAPHY

[1] Alessandro Achille. "Emergent Properties of Deep Neural Networks". PhD thesis. UCLA, 2019. URL: https://escholarship.org/uc/item/8gb8x6w9.

[2] Alessandro Achille, Glen Mbeng, and Stefano Soatto. *Dynamics and Reachability of Learning Tasks*. 2018. arXiv: 1810.02440 [cs.LG].

[3] Alessandro Achille, Matteo Rovere, and Stefano Soatto. *Critical Learning Periods in Deep Neural Networks*. 2017. arXiv: 1711.08856 [cs.LG].

[4] Alessandro Achille and Stefano Soatto. "Emergence of Invariance and Disentangling in Deep Representations". In: *J. Mach. Learn. Res.* 19.1 (Jan. 2018), pp. 1947–1980. ISSN: 1532-4435.

[5] Alessandro Achille and Stefano Soatto. "Information Dropout: Learning Optimal Representations Through Noisy Computation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.12 (2018), pp. 2897–2905.

[6] Alessandro Achille and Stefano Soatto. *Where is the Information in a Deep Neural Network?* 2019. arXiv: 1905.12213 [cs.LG].

[7] Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. *The Information Complexity of Learning Tasks, their Structure and their Distance*. 2019. arXiv: 1904.03292 [cs.LG].

[8] O. Aftab, A. Kim Cheung, S. Thakkar, and N. Yeddanapudi. *Information Theory: Information Theory and the Digital Age*. Web document for 6.933 Project History, Massachusetts Institute of Technology. 2001. URL: http://web.mit.edu/6.933/www/Fall2001/Shannon2.pdf.

[9] Aristotle. *Aristotle: Nicomachean Ethics*. Cambridge Texts in the History of Philosophy. Cambridge University Press, 2000. DOI: 10.1017/CBO9780511802058.

[10] Jonathan Baxter. "A model of inductive bias learning". In: *Journal of artificial intelligence research* 12 (2000), pp. 149–198.

[11] Yoshua Bengio. "Deep learning of representations for unsupervised and transfer learning". In: *Proceedings of ICML workshop on unsupervised and transfer learning*. 2012, pp. 17–36.

[12] Avrim Blum. *Machine learning theory*. Tech. rep. Carnegie Mellon University, School of Computer Science, 2007. URL: https://www.cs.cmu.edu/~avrim/Talks/Talks/mlt.pdf.

[13] Ariel Caticha. *Lectures on Probability, Entropy, and Statistical-Physics*. arXiv: 0808.0012 [physics.data-an].

[14] T. M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd ed. OCLC: ocm59879802. Wiley-Interscience, 2006. ISBN: 9780-471-2419-5-9.

[15] Daniel Dennett. "Darwin's "strange inversion of reasoning"". In: vol. 106. Supplement 1. National Academy of Sciences, 2009, pp. 10061–10065. DOI: 10.1073/pnas.0904433106. eprint: https://www.pnas.org/content/106/Supplement_1/10061.full.pdf. URL: https://www.pnas.org/content/106/Supplement_1/10061.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. arXiv: 1810.04805 [cs.CL].

[17] Richard Feynman. *The Character of Physical Law*. Modern Library, 1994. ISBN: 0-679-60127-9.

[18] Martin Gardner. *Logic machines and diagrams*. McGraw-Hill Book Company, 1959.

[19] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN: 9780-262-0356-1-3.

[20] Fred Guth. *An Information Theoretical Transferability Metric*. Tech. rep. UnB, June 2019.

[21] Fred Guth and Teofilo Emidio de Campos. *Research Frontiers in Transfer Learning – a systematic and bibliometric review*. 2019. arXiv: 1912.08812 [cs.DL].

[22] Aurélien Géron. *A Short Introduction to Entropy, Cross-Entropy and KL-Divergence*. [Online; Last accessed on 2020-03-08.] Feb. 5, 2018. URL: https://www.youtube.com/watch?v=ErfnhcEV1O8.

[23] David Haussler. "Quantifying inductive bias: AI learning algorithms and Valiant's learning framework". In: *Artificial intelligence* 36.2 (1988), pp. 177–221.

[24] Jeremy Howard. *Intro to Machine Learning: Lesson 1*. [Online; Last accessed on 2020-04-20. Speech at 38'20" in the video.] 2018.

[25] Jeremy Howard and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *ACL*. Association for Computational Linguistics, 2018. URL: http://arxiv.org/abs/1801.06146.

[26] Bryce Huebner. *The Philosophy of Daniel Dennett*. Oxford University Press, 2017.

[27] David Hume. *Tratado da natureza humana-2a Edição*. Editora UNESP, 2009. ISBN: 97885-7139-901-3.

[28] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. ISBN: 0-521-59271-2.

[29] Sham Kakade and Ambuj Tewari. *VC Dimension of Multilayer Neural Networks, Range Queries*. [Online; last accessed on February 3rd, 2020]. 2008. URL: https://ttic.uchicago.edu/~tewari/lectures/lecture12.pdf.

[30] Daniel Klein. *Mighty mouse*. [Online; Published: 2018-12-19. Accessed: 2020-01-16]. 2018. URL: https://www.technologyreview.com/s/612529/mighty-mouse/.

[31] Shane Legg and Marcus Hutter. *A Collection of Definitions of Intelligence*. 2007. arXiv: 0706.3639 [cs.AI].

[32] Henry W Lin, Max Tegmark, and David Rolnick. "Why does deep and cheap learning work so well?" In: *Journal of Statistical Physics* 168.6 (2017), pp. 1223–1247.

[33] Zachary C. Lipton and Jacob Steinhardt. *Troubling Trends in Machine Learning Scholarship*. 2018. arXiv: 1807.03341 [stat.ML].

[34] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. USA: Cambridge University Press, 2002. ISBN: 0521642981.

[35] Robert Beverley MacKenzie. *The Darwinian Theory of the Transmutation of Species Examined*. J. Nisbet, 1868, p. 318.

[36] Lisa Margonelli. *Collective Mind in the Mound: How Do Termites Build Their Huge Structures?* [Online; Last accessed: 2020-04-26]. Apr. 2016. URL: https://www.nationalgeographic.com/news/2014/8/140731-termites-mounds-insects-entomology-science/.

[37] Adrienne Mayor. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton University Press, 2018. ISBN: 9780-691-18351-0.

[38] Warren S. McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[39] Brockway McMillan. "The Basic Theorems of Information Theory". In: *Ann. Math. Statist.* 24.2 (June 1953), pp. 196–219. DOI: 10.1214/aoms/1177729028. URL: https://doi.org/10.1214/aoms/1177729028.

[40] Rodrigo F. Mello. *Statistical Learning Theory*. [Online;Published: 2018-10-03. Last Accessed: 2020-04-22]. Oct. 2018. URL: https://www.youtube.com/watch?v=KTrRap4Spd0.

[41] Rodrigo F. Mello and Moacir Antonelli Ponti. *Machine learning: a practical approach on the statistical learning theory*. Springer, 2018.

[42] Tom M. Mitchell. *The Need for Biases in Learning Generalizations*. Tech. rep. Rutgers University, 1980.

[43] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN: 9780-262-01825-8.

[44] John R. Pierce. *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications. ISBN: 0486240614.

[45] Stuart J. Russell, Peter Norvig, and Ernest Davis. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2010. ISBN: 9780-13-60425-9-4.

[46] Andrew M. Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D. Tracey, and David D. Cox. "On the information bottleneck theory of deep learning". In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019), p. 124020.

[47] John G. Saxe. *The blind men and the elephant*. Enrich Spot Limited, 2016.

[48] Claude E. Shannon. "A mathematical theory of communication". In: *Bell system technical journal* 27.3 (1948), pp. 379–423.

[49] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949. ISBN: 978-0-252-72548-7.

[50] John Shawe-Taylor and Omar Rivasplata. *Statistical Learning Theory - a Hitchhiker's Guide (NeurIPS 2018)*. [Online;Published: 2018-12-09. Last Accessed: 2020-04-22]. Dec. 2018. URL: https://www.youtube.com/watch?v=m8PLzDmW-TY&t=780s.

[51] Ravid Shwartz-Ziv and Naftali Tishby. "Opening the Black Box of Deep Neural Networks via Information". In: (2017). arXiv: 1703.00810 [cs.LG].

[52] Ravid Shwartz-Ziv and Naftali Tishby. *Representation Compression and Generalization in Deep Neural Networks*. 2019. URL: https://openreview.net/forum?id=SkeL6sCqK7.

[53] Noam Slonim. "The information bottleneck: Theory and applications". PhD thesis. Hebrew University, 2002.

[54] Leslie N. Smith and Nicholay Topin. "Super-convergence: Very fast training of neural networks using large learning rates". In: *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*. Vol. 11006. International Society for Optics and Photonics. 2019, p. 1100612.

[55] Jimmy Soni and Rob Goodman. *A mind at play: how Claude Shannon invented the information age*. Simon and Schuster, 2017.

[56] Damian Radoslaw Sowinski. "Complexity and stability for epistemic agents: The foundations and phenomenology of configurational Entropy". PhD thesis. 2016.

[57] James V. Stone. *Information theory: a tutorial introduction*. Sebtel Press, 2015.

[58] Alexander Terenin and David Draper. "Cox's Theorem and the Jaynesian Interpretation of Probability". In: (2015). arXiv: 1507.06597 [math.ST].

[59] Naftali Tishby. *Information Theory of Deep Learning*. [Online; Published: 2017-10-16. Last Accessed: 2020-03-06]. Oct. 16, 2017. URL: https://www.youtube.com/watch?v=FSfN2K3tnJU.

[60] Naftali Tishby and Noga Zaslavsky. "Deep learning and the information bottleneck principle". In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE. 2015, pp. 1–5.

[61] L. G. Valiant. "A theory of the learnable". In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing - 84*. ACM Press, 1984. DOI: 10.1145/800057.808710.

[62] Ulrike Von Luxburg and Bernhard Schölkopf. "Statistical learning theory: Models, concepts, and results". In: *Handbook of the History of Logic*. Vol. 10. Elsevier, 2011, pp. 651–706.

[63] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[64] Torsten N. Wiesel. "Postnatal Development of the Visual Cortex and the Influence of Environment". In: *Nature* 299.5884 (Oct. 1982), pp. 583–591. ISSN: 1476-4687. DOI: 10.1038/299583a0.

[65] David H. Wolpert and William G. Macready. "No free lunch theorems for optimization". In: *IEEE transactions on evolutionary computation* 1.1 (1997), pp. 67–82.

[66] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. "Efficient compression in color naming and its evolution". In: vol. 115. 31. National Academy of Sciences, 2018, pp. 7937–7942.

[67] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. *Understanding deep learning requires rethinking generalization*. 2016. arXiv: 1611.03530 [cs.LG].

[68] Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. "Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach". In: *International Conference on Learning Representations*. 2019.