Universidade de Brasília

The emergence of an Information Bottleneck Theory *of* Deep Learning

Dissertation presented for the conclusion of the Master Degree in Computer Science

Frederico Guth

December 2021

FICHA CATALOGRÁFICA DE TESES E DISSERTAÇÕES

Está página existe apenas para indicar onde a ficha catalográfica gerada para dissertações de mestrado e teses de doutorado defendidas na UnB. A Biblioteca Central é responsável pela ficha, mais informações nos sítios:

1. http://www.bce.unb.br
2. http://www.bce.unb.br/elaboracao-de-fichas-catalograficas-de-teses-e-dissertacoes

Esta página não deve ser inclusa na versão final do texto.

Universidade de Brasília
Institute of Exact Sciences
Department of Computer Science

The emergence of an Information Bottleneck Theory *of* Deep Learning

Frederico Guth

Dissertation presented for the conclusion of the Master Degree in Computer Science

Prof. Teófilo Emidio de Campos (Supervisor)
Universidade de Brasília

Prof. Shawe-Taylor
University College London

Prof. Moacir Ponti
Universidade de São Paulo

Prof. Ricardo Jacobi
Computer Science Graduate Program Coordinator

Brasília, December 1st, 2021

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. In the past, the false solution was to "add computation power on problems", today we try "piling 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. It is a known fact, however, that learning from way fewer samples is possible: humans show a much better generalisation ability than the current state of the art artificial intelligence. To achieve such a feat, a better understanding of how generalisation works is needed, in particular in deep neural networks. However, the practice of modern machine learning has outpaced its theoretical development. In particular, *"traditional measures of model complexity struggle to explain the generalization ability of large artificial neural networks"* [Zha+16]. There is yet no established new general theory of learning which handles this pseudo-paradox. In 2015, Naftali Tishby and Noga Zaslavsky published a seminal theory of learning based on the information-theoretical concept of the bottleneck principle with the potential of filling this gap. This dissertation aims to investigate the efforts using the information bottleneck principle to explain the generalisation capabilities of deep neural networks, consolidate them into a comprehensive digest and analyse its relation to current machine learning theory.

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization.*

Resumo Extendido

Na última década, assistimos estupefatos uma miríade de sucessos em Aprendizado Profundo (Deep Learning (DL)). 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 bases de dados de treinamento entre grandes corporações, suscitando preocupações sobre privacidade e concentração de poder. É fato, entretanto, que aprender com muito menos dados é possível: humanos demonstram uma habilidade de generalização muito superior ao estado-da-arte atual em Inteligência Artificial.

Para atingir tal capacidade, precisamos entender melhor como o aprendizado ocorre em Deep Learning. A prática tem se desenvolvido mais rapidamente que a teoria na área. Em particular, Zhang et al. demonstraram que modelos de deep learning são capazes de memorizar rótulos aleatórios, ainda assim apresentam alto poder de generalização [Zha+16]. A atual teoria de aprendizado de máquinas não explica tal poder de generalização em modelos superparametrizados.

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization.*

Em 2015, Naftali Tishby e Noga Zaslavsky publicaram uma teoria de aprendizado baseado no princípio do gargalo de informação (information bottleneck) [TZ15a]. Tal teoria sucitou interesse e desconfiança pela academia, tendo vários de seus artigos primordiais sido contestados em artigos posteriores. Esta dissertação 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 deste novo desenvolvimento teórico denominado Teoria do Gargalo de Informação (Information Bottleneck Theory (IBT)) que mostre seus pontos fortes e fracos e oportunidades de pesquisa.

[TZ15a] Tishby and Zaslavsky, 'Deep learning and the information bottleneck principle'.

A BUSCA DOS FUNDAMENTOS

Nesta investigação, partimos de uma discussão filosófica sobre o que é inteligência e o que significa aprender (Capítulo 2) e, passo a passo (Capítulos 3 e 5), mostramos em que fundamentos a teoria vingente de aprendizado de máquinas (Machine Learning Theory (MLT)), assim

como a emergente (Information Bottleneck Theory (IBT)) se apoiam. Pudemos assim perceber que ambas teorias se baseiam em um conjunto muito similar de premissas. A maior diferença é que Information Bottleneck Theory (IBT) assume o uso de variáveis aleatórias discretas de espaços finitos. Entretanto, tal limitação não é significativa, uma vez que pesquisas já demonstraram que é possível tornar o erro de quantização arbitrariamente pequeno conquanto haja memória para tanto [Ris86; HVC93]. Além disso, Information Bottleneck Theory (IBT) não invalida nenhum resultado de Machine Learning Theory (MLT), pelo contrário, apresenta uma nova narrativa que nos permite conciliar os resultados teóricos com os fenômenos observados, quando medimos complexidade como a quantidade de informação nos pesos de um modelo, e não a sua quantidade de parâmetros.

[Ris86] Rissanen, 'Stochastic complexity and modeling'.

[HVC93] Hinton and Van Camp, 'Keeping the neural networks simple by minimizing the description length of the weights'.

Essa investigação nos permitiu sintetizar o desenvolvimento teórico em Teoria da Informação (Information Theory (IT)) e Machine Learning Theory (MLT) em uma abodagem que denominamos PAC-Shannon (capítulo 6) em que partimos dos teoremas fundamentais de Shannon em Information Theory (IT) e provamos limites para erro de generalização em aprendizado.

EXPLICANDO A NOVA TEORIA

Tishby propôs que vejamos aprendizado como um problema de codificação (Capítulo 7). Nessa perspectiva, os dados de entrada contém informação de um alvo, uma variável rótulo, a qual não temos acesso; o problema de aprendizado é encontrar o codificador-decodificador que explique nossos nossos dados de treinamento; o conjunto de dados (*dataset*) de treinamento é a definição da tarefa (padronagem estrutural dos dados) que se quer aprender. Em Information Bottleneck Theory (IBT), generalização não depende do espaço de hipóteses do modelo, mas apenas dos limites de compressibilidade do *dataset*. Limites esses definidos pelos teoremas de Shannon (Capítulo 5). Enquanto Machine Learning Theory (MLT) é agnóstico à distribuição dos dados e modelo-dependente, Information Bottleneck Theory (IBT) é agnóstico ao modelo e distribuição-dependente. Esta perspectiva, se relaciona perfeitamente com a teoria algorítimica da informação (Kolmogorov-Chaitin complexity (KC)) (Seção 5.8.1).

Essa visão de informação como medida de complexidade, nos permite analisar o treinamento enquanto ele acontece. Ou seja, para aqueles que se sentem desconfotáveis com o fato da teoria corrente ver modelos como uma caixa-preta, onde só se analisa a entrada e a saída, medidas de informação nos permitem entender o que ocorre durante

o treinamento. Essa análise leva a surpreendente conclusão de que o aprendizado tem duas fases distintas: uma fase de ajuste e outra de compressão. Primeiro, na fase de ajuste, o modelo memoriza os dados, minimizando rapidamente o erro e usando muita informação que é peculiar apenas ao *dataset* utilizado e não à variável-alvo; na fase posterior de compressão, o modelo tenta esquecer o máximo possível sobre os dados de entrada enquanto mantém a informação sobre o alvo, reduzindo a quantidade de informação no modelo.

PONTOS FORTES E FRACOS E DE OPORTUNIDADE EM IBT

Partindo do princípio do gargalo de Teoria da Informação demonstramos a coesão interna desta narrativa alternativa (Capítulo 8), e mostramos o embasamento teórico de práticas em Deep Learning (DL), como o uso de Entropia Cruzada como função custo na otimização de modelos; e seus fenômenos, como a generalização de modelos superparametrizados e períodos críticos de aprendizado [ARS17](Capítulo 9).

[ARS17] Achille et al., *Critical Learning Periods in Deep Neural Networks.*

A Information Bottleneck Theory (IBT), entretanto, está longe de ser um desenvolvimento teórico completo. Falta de rigor, definição e objetivos claros em alguns dos seus artigos científicos primeiros deram razão ao ceticismo e até discrédito em que a teoria passou a ser vista. O trabalho de Achille e Soatto (Capítulos 8 e 9) foi menos ambicioso em suas alegações e mais rigoroso, resolvendo alguns dos problemas da apresentação inicial da teoria, mas não se propõe a ser completo. A presente dissertação também presta a esse papel de dar um pouco mais de rigor e clareza aos princípios assumidos, mas há ainda muito o que se desenvolver:

formulação PAC: Seria possível criar uma formulação PAC que dependa apenas de β, uma vez que esse parâmetro representa um único limite (ϵ, δ).

Novas estratégias de otimização: Se o treinamento tem duas fases como preconiza Information Bottleneck Theory (IBT), isso nos permite usar estratégias de otimização diferenciadas para cada uma.

Transferência de Aprendizado: Se, em Information Bottleneck Theory (IBT), complexidade depende apenas da compressibilidade do *dataset* e de um nível desejado de performance e generalização (β), podemos analisar a complexidade de datasets e montar uma topologia de tarefas com a predição da similar-

iedade (distância) entre *datasets* e relacionar tais resultados teóricos com resultados empíricos como os obtidos por Zamir et al. [Zam+18].

[Zam+18] Zamir et al., 'Taskonomy: Disentangling task transfer learning'.

Processos ergódicos: Os princípios de teoria da informação não requerem amostragem independentes e identicamente distribuídas, mas apenas que sejam processos ergódicos.

Conexão com mecânica estatística: A área de Mecânica Estatística já se desenvolve em Física há mais de um século. A conexão de aprendizado de máquina com teoria da informação permite a exploração de resultados nessa área de Física (como fizeram [CS18; Cha+19a]).

[CS18] Chaudhari and Soatto, 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'.

[Cha+19a] Chaudhari et al., 'Entropy-SGD: Biasing gradient descent into wide valleys'.

Em resumo, a presente dissertação foi capaz de estabelecer que Information Bottleneck Theory (IBT) está longe de ser uma teoria rigorosa e completa, mas que é uma interessante teoria emergente que apresenta ainda muitas oportunidades de pesquisa e merece atenção.

Contents

Notation

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

NUMBERS AND ARRAYS

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

INDEXING

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

LINEAR ALGEBRA OPERATIONS

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

CALCULUS

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

SETS

$\mathbb{A}$	A set
$\wp(\mathrm{A})$	The powerset (the set of subsets) of A
$\{0,1\}^n$	The set containing n 0 or 1s
$\{0,\ldots,n\}$	The set of all integers between 0 and n
$a \in \mathbb{A}$	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 $\mathbb{A}$

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
$\mathbb{S}$	A sample, *i.e.* a set of training examples
x_i	The i-th example (input) from a dataset
$\mathbf{W}^{(i)}$	The matrix $\mathbf{W}$ of weights in the i-th layer of a network
y_i	The target associated with $\boldsymbol{x}_i$ for supervised learning

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, $f(g(\cdot))$
$f(\boldsymbol{x}; \boldsymbol{\theta}) \equiv f_{\boldsymbol{\theta}}(\boldsymbol{x})$	A function of $\boldsymbol{x}$ parametrised 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 nonlinear activation function
x^+	Positive part of x, i.e., $\max(0, x)$
$\mathbb{1}_{[condition]}$	is the **indicator function** and is 1 if the condition is true, 0 otherwise

PROBABILITY THEORY

Ω	A experiment or sample space
ω	An outcome (an example)
A	An event
$\mathrm{A} \perp \mathrm{B}$	The events A and B are independent
X	A random variable
$X \perp Y$	The random variables X and Y are independent
$\mathrm{P}(\mathrm{A} \mid \mathrm{B})$	The probability of an event A given the event B happened
$\mathrm{P}(X = a_i) \equiv \mathrm{P}_X \equiv p(a_i) \equiv p_i \equiv p$	A probability distribution over a random variable (discrete or continuous defined by the context)
$a \sim p$	An example a drawn from distribution p
$\mathbb{E}_{X \sim p}[x] \equiv \mathbb{E}_p X \equiv \langle X \rangle_p$	Expectation of x w.r.t. $p(x)$, i.e. $\sum_{i=1}^{k} x_i \, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$
$\sigma^2(f(x))$	Variance of $f(x)$ under $p(x)$
$\mathcal{N}(\boldsymbol{x}; \mu, \sigma^2)$	Gaussian distribution over $\boldsymbol{x}$ with mean μ and variance σ^2

INFORMATION THEORY

$\mathrm{H}[X]$	The entropy of a random process X (bits)
$\mathrm{D}_{\mathrm{KL}}(p\|q)$	Kullback-Leibler divergence of distribution p and q
$\mathrm{H}[X\|Y]$	The conditional entropy of a random process X given Y. (bits)
$\mathrm{R}[X] \equiv Я[X]$	The rate of a transmission of X (bits)
$\mathrm{H}_{p,q}[\mathrm{X}]$	The cross-entropy of X between its true distribution p and a modelled distribution q (bits)
$\mathrm{C}[X;Y]$	The capacity of a channel between X and Y (bits)
$\mathrm{I}[X;Y]$	The mutual information between X and Y (bits)
$\mathbb{T}_\delta(X) \equiv \mathbb{T}_\epsilon(X) \equiv \mathbb{A}_X^\delta$	The typical set of X

Acronyms

AEP Asymptotic Equipartition Property 81

AI Artificial Intelligence 2

ANN Artificial Neural Network 18

CV Computer Vision 24

DL Deep Learning 9

DNN Deep Neural Network 26

DPI Data Processing Inequality 86

DVIB Deep Variational Information Bottleneck 148

ELBO Evidence Lower Bound 134

FIM Fisher Information Matrix 151

GPU Graphical Processor Unit 25

IB Information Bottleneck 121

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

In his acceptance speech for the Test-of-Time award in NeurIPS 2017,[1] Ali Rahimi[2] started a controversy by frankly declaring *'Machine learning has become alchemy'* [Rah18, 12'10"]. His concerns on the lack of theoretical understanding of machine learning for critical decision-making are rightful: *'We are building systems that govern healthcare and mediate our civic dialogue. We would influence elections. I would like to live in a society whose systems are built on top of verifiable, rigorous, thorough knowledge and not on alchemy.'*

The next day, Yann LeCun[3] responded: *'Criticising an entire community (. . .) for practising "alchemy", simply because our current theoretical tools have not caught up with our practice is dangerous.'*

Both researchers, at least, agree upon one thing: the practice of machine learning has outpaced its theoretical development. That is certainly a research opportunity.

1.1.1 *A Tale of Babylonians and Greeks*

Richard Feynman (Figure 1.1) used to lecture this story [Fey94]: Babylonians were pioneers in mathematics; Yet, the Greeks took the credit. We are used to the Greek way of doing Math: start from the most basic axioms and build up a knowledge system. 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 as science advances, one can worry about reformulation, simplification and

[1] Conference on Neural Information Processing.

[2] Research Scientist, Google.

[Rah18] Rahimi, *Ali Rahimi NIPS 2017 Test-of-Time Award Presentation Speech*. URL: https://youtu.be/x7psGHgatGM

[3] Deep Learning pioneer and Turing award winner (2018). https://www.facebook.com/yann.lecun/posts/10154938130592143

FIGURE 1.1: Richard Feynman, Nobel laureate physicist.[4]

[4] Except when otherwise stated, all images were created by the author.

[Fey94] Feynman, *The Character of Physical Law*.

ordering. Scientists are Babylonians; mathematicians are Greeks.

Mathematics and science are both tools for knowledge acquisition. They are also social constructs that rely on peer-reviewing. They are somewhat 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, only one experiment was needed 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. Mathematics is a 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.

In science, there are no axioms: a falsifiable hypothesis/theory is proposed, and logical conclusions (predictions) from the theory are empirically tested. Despite inferring hypotheses by induction, there is no influence of psychology in the process. A tested hypothesis is not absolute truth. A hypothesis is never verified, only falsified by experiments [Pop04, p. 31-50]. Scientific knowledge is belief justified by experience; there are degrees of plausibility.

[Pop04] Popper, *A Lógica da Pesquisa Científica*.

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.

1.1.2 *The importance of theoretical narratives*

Science is a narrative of how we understand Nature [GS18]. In science, we collect facts, but they need interpretation. The logical conclusion from the hypothesis that predicts some behaviour in nature gives a plausible *meaning* to what we observed.

[GS18] Gleiser and Sowinski, 'The Map and the Territory'.

To illustrate, take the ancient human desire of flying. There have always 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) [Far16]. While concepts like lift, stability, and control were poorly understood, most human flight attempts ended in severe injury or even death. It did not matter how much evidence, how many hours of seeing different animals flying, those ludicrous brave men

[Far16] Farrington, *The blitzed city : the destruction of Coventry, 1940*.

experienced; the *meaning* they took from what they saw was wrong, and their predictions incorrect.

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

They did not die in vain[5]; 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 observations in animal flight from the aerodynamics perspective, we learned to fly better than any other animal before. Science works by a "natural selection" of ideas, where only the fittest ones survive until a better one is born. Chaitin also points out that an idea has "fertility" to the extent to which it "illuminates us, inspires us with other ideas, and suggests unsuspected connections and new viewpoints" [Cha06, p. 9].

[5] Those "researchers" deserved, at least, a Darwin Award of Science. The Darwin Award is satirical honours that recognise individuals who have unwillingly contributed to human evolution by selecting themselves out of the gene pool.

[Cha06] Chaitin, *Meta Math! The Quest for Omega*.

Being a Babylonian enterprise, science has no clear path. One of the exciting facts one can learn by studying its history is that robust discoveries have arisen through the study of phenomena in human-made devices [Pie]. For instance, Carnot's first and only scientific work [Kle74] gave birth to thermodynamics: the study of energy, the conversion between its different forms, and the ability of energy to do work; *i.e.* the science that explains how steam engines work. However, steam engines came before Carnot's work and were studied by him. Such human-made devices may present a simplified instance of more complex natural phenomena.

[Pie] Pierce, *An Introduction to Information Theory: Symbols, Signals and Noise.*

[Kle74] Klein, 'Carnot's contribution to thermodynamics'.

Another example is Information Theory. Several insights of Shannon's theory of communication were generalisations of ideas already present in Telegraphy [Sha48]. New theories in artificial intelligence

[Sha48] Shannon, 'A mathematical theory of communication'.

can, therefore, be developed from insights in the study of deep learning phenomena.[6]

[6] Understanding human intelligence using artificial intelligence is a field of study called Computational Neuroscience.

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, 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, despite its undeniable usefulness:

- Even among AI researchers, there is a trend of "mathiness" and speculation disguised as explanations in conference papers [LS18].
- There are few venues for papers that 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.

[LS18] Lipton and Steinhardt, *Troubling Trends in Machine Learning Scholarship*.

While physicists rejoice in finding phenomena that contradict current theories, computer scientists get baffled. In Natural Sciences, unexplained phenomena lead to theoretical development. Some believe they bring "winters", periods of progress stagnation and lack of funding in AI.[7]

[7] This seems to be Yann LeCun's opinion: '*Why [Rahimi's position is] dangerous? It is exactly this kind of attitude that lead the ML community to abandon neural nets for over 10 years, despite ample empirical evidence that they worked very well in many situations.*' However, due to all possible alternative explanations (lack of computational power, no availability of massive annotated datasets), it seems harsh or simply wrong to blame theorists.

Artificial Intelligence has been through several of the aforementioned "winters". In 1957, Herbert Simon[8] famously predicted that within ten years, a computer would be a chess champion [RND10, 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.

[8] Herbert Simon (1916–2001) received the Turing Award in 1975, and the Nobel Prize in Economics in 1978.

[RND10] Russell et al., *Artificial Intelligence*.

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 (besides the number of steps). The matter of concern is that it currently predicts that generalisation requires simpler models in terms of parameters. In total disregard to the theory, deep learning models have shown spectacular generalisation power with hundreds of millions of parameters (and even more impressive overfitting capacity [Zha+16]).

1.2 PROBLEM

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 "add computation power" on problems, today we try to solve them by "piling data"(Figure 1.3). Such behaviour has triggered a winner-takes-all competition for who owns more data (ourdata) amidst a handful of large corporations, raising ethical concerns about privacy and concentration of power [O'N16].

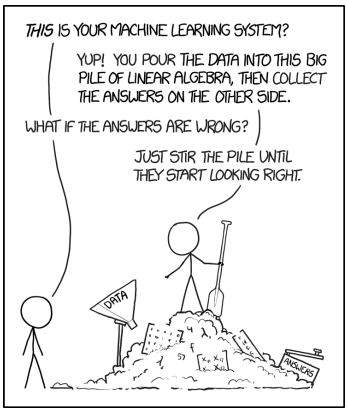


FIGURE 1.3: Source: https://xkcd.com/1838/. Reprinted with permission.

Nevertheless, we know 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 might reshape the foundations of machine learning theory [Zha+16].

[O'N16] O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*.

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization*.

1.2.1 *Possible new explanation in the horizon*

In 2015, Tishby and Zaslavsky proposed a theory of deep learning [TZ15b] based on the information-theoretical concept of the bottleneck principle, of which Tishby is one of the authors. Later, in 2017, Shwartz-Ziv and Tishby followed up on the Information Bottleneck Theory (IBT) with the paper 'Opening the Black Box of Deep Neural Networks via Information', which was presented in a well-attended workshop[9], with appealing visuals that clearly showed a *"phase transition"* happening during training. The video posted on Youtube [Tis17a] became a "sensation"[10], and received a wealth of publicity when well-known researchers like Geoffrey Hinton[11], Samy Bengio (Apple) and Alex Alemi (Google Research) have expressed interest in Tishby's ideas [Wol17]. they are called formal languages.

> *I believe that the information bottleneck idea could be very important in future deep neural network research.*
>
> — Alex Alemi

Andrew Saxe (Harvard University) rebutted Shwartz-Ziv and Tishby claims in 'On the Information Bottleneck Theory of Deep Learning' and was followed by other critics. According to Saxe, it was impossible to reproduce [ST17]'s experiments with different parameters.

[TZ15b] Tishby and Zaslavsky, 'Deep learning and the information bottleneck principle'.

[9] Deep Learning: Theory, Algorithms, and Applications. Berlin, June 2017 http://doc.ml.tu-berlin.de/dlworkshop2017

[Tis17a] Tishby, *Information Theory of Deep Learning*.
URL: https://youtu.be/bLqJHjXihK8

[10] By the time of this writing, this video as more than 84k views, which is remarkable for an hour-long workshop presentation in an academic niche. https://youtu.be/bLqJHjXihK8

[11] Another Deep Learning Pioneer **and** Turing award winner (2018).

[Wol17] Wolchover, *New Theory Cracks Open the Black Box of Deep Learning*.

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

Has the initial enthusiasm on the IBT been unfounded? Have we let us "fool ourselves" by beautiful charts and a good story?

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 the current machine learning theory. There is yet no established new general theory of learning which handles this problem.

IBT was proposed as a possible new theory with the **potential** of filling the theory-practice gap. Unfortunately, to the extent of our knowledge, **there is still no comprehensive digest of IBT nor an analysis of how it relates to current Machine Learning Theory (MLT).**

1.3 OBJECTIVE

This dissertation aims to investigate *to what extent* can the emergent Information Bottleneck Theory help us better understand Deep Learning and its phenomena, especially generalisation, presenting its strengths, weaknesses and research opportunities.

1.3.1 *Research Questions*

1. What are the fundamentals of IBT? How do they differ from the ones from MLT?
2. What is the relationship between IBT and current MLT? How different or similar they are?
3. Is IBT capable of explaining the phenomena MLT already explains?
4. Does IBT invalidate results in MLT?
5. Is IBT capable of explaining phenomena still not well understood by MLT?
6. What are Information Bottleneck Theory's (IBT) strengths?
7. What are Information Bottleneck Theory's (IBT) weaknesses?
8. What has been already developed in IBT?
9. What are Information Bottleneck Theory's (IBT) research opportunities?

1.4 METHODOLOGY

Scope: Given that IBT is yet not a well-established learning theory, there were two difficulties that the research had to address:

a) There is a growing interest in the subject, and new papers are published every day. It was essential to select literature and restrain the analysis.

b) Early on, the marks of an emergent theory in its infancy manifested in the form of missing assumptions, inconsistent notation, borrowed jargon, and seeming missing steps. Foremost, it was unclear what was missing from the theory and what was missing in our understanding.

An initial literature review on IBT was conducted to define the scope.[12] We then chose to narrow the research to **Information Bottleneck Theory's (IBT) theoretical perspective on generalisation**, where we considered that it could bring fundamental advances. We made the deliberate choice of going deeper in a limited area of IBT and not broad, leaving out a deeper experimental and application analysis, all the work on Information-Theoretic Learning (ITL)[13] [Pri10] and statistical-mechanics-based analysis of SGD [CS18; Cha+19b]. From this set of constraints, we chose a list of pieces of IBT literature to go deeper (Appendix A).

[12] Not even the term IBT is universally adopted.

[13] ITL makes the opposite path we are taking, bringing concepts of machine learning to information theory problems.

[Pri10] Principe, *Information theoretic learning: Renyi's entropy and kernel perspectives*.

[CS18] Chaudhari and Soatto, 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'.

[Cha+19b] Chaudhari et al., 'Entropy-sgd: Biasing gradient descent into wide valleys'.

Background analysis: In order to answer research questions 1 to 4, we discuss the epistemology of Artificial Intelligence to choose fundamental axioms (definition of intelligence and the definition of knowledge) with which we deduced from the ground up MLT, Information Theory (IT) and IBT, revealing hidden assumptions, pointing out similarities and differences. By doing that, we built a "genealogy" of these research fields. This comparative study was essential for identifying missing gaps and research opportunities.

IBT literature digest: In order to answer research questions 5 to 9, we first dissected the selected literature (Appendix A) and organised scattered topics in a comprehensive sequence of subjects.

IBT analysis: In the process of the literature digest, we identified results, strengths, weaknesses and research opportunities.

1.5 CONTRIBUTIONS

In the research conducted, we produced three main results that, to the extent of our knowledge, are original:

IBT Digest and Analysis: The dissertation itself is the main expected result: a comprehensive digest of the IBT literature and a snapshot analysis of the field in its current form, focusing on its theoretical implications for generalisation.

PAC-Shannon: We propose an Information-Theoretical learning problem different from Minimum Description Length (MDL) proposed by [HVC93], but with the same results. We derived PAC bounds for this proposed learning problem using Shannon's Theorems 6.3 to 6.6.

Layers reduce the effective hypothesis space: We believe we found a logical misstep in [Ach19] explanation for the role of layers in Deep Representation in the IBT perspective. We then fix the explanation by proving the counter-intuitive argument that layers reduce the model's "effective" hypothesis space (Section 9.5.2).

[Ach19] Achille, 'Emergent Properties of Deep Neural Networks'.
URL: https://escholarship.org/uc/item/8gb8x6w9

1.6 DISSERTATION PREVIEW AND OUTLINE

The dissertation is divided into two parts (Part I and Part III), with a break in the middle (Part II).

1. Background (Part I)
   - 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 machine learning theory.
   - Chapter 3 — Probability Theory: The chapter derives propositional calculus and probability theory from a list of desired characteristics for epistemic agents. It also presents basic Probability Theory concepts.
   - Chapter 4 — Machine Learning Theory: The chapter presents the theoretical framework of Machine Learning, the PAC model, theoretical guarantees for generalisation, and expose its weaknesses concerning Deep Learning phenomena.
   - Chapter 5 — Information Theory: The chapter derives Shannon Information from Probability Theory, explicates some implicit assumptions, and explains basic Information Theory concepts.
2. Intermezzo (Part II)

- Chapter 6 — Information-Theoretical Epistemology: This chapter closes the background part and opens the IBT part of the dissertation. It shows the connection of IT and MLT in the learning problem, proves that Shannon theorems can be used to prove PAC bounds and present the Minimum Description Length (MDL) Principle, an earlier example of this kind of connection.

3. The emergence of a theory (Part III)

   - Chapter 7 — IB Principle: Explains the IB method and its tools: Kullback-Leibler divergence (D_{KL}) as a natural distortion (loss) measure, the IB Lagrangian and the Information Plane.
   - Chapter 8 — IB and Representation Learning: Presents the learning problem in the IBT perspective (not specific to Deep Learning (DL)). It shows how some usual choices of the practice of DL emerge naturally from a list of desired properties of representations. It also shows that the information in the weights bounds the information in the activations.
   - Chapter 9 — IB and Deep Learning: This chapter presents the IBT perspective specific to Deep Learning. It presents IBT analysis of Deep Learning training, some examples of applications of IBT to improve or create algorithms; and the IBT learning theory of Deep Learning. We also explain Deep Learning phenomena in the IBT perspective.
   - Chapter 10 — Conclusion: In this chapter, we present a summary of the findings, answer the research questions, and present suggestions for future work.

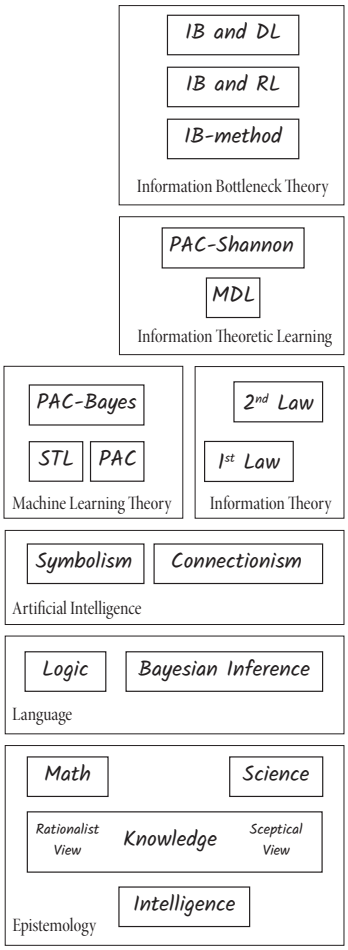


FIGURE 1.4: IBT "genealogy" tree.

We found out that IBT does not invalidate MLT; it just interprets complexity not as a function of the data (number of parameters) but as a function of the information contained in the data. With this interpretation, there is no paradox in improving generalisation by adding layers.

Furthermore, they both share more or less the same "genealogy" of assumptions. IBT can be seen as particular case of MLT. Nevertheless, IBT allows us to better understand the training process and provide a different narrative that helps us comprehend Deep Learning phenomena in a more general way.

Part I

BACKGROUND

2

Artificial Intelligence

'I visualise a time when we will be to robots what dogs are to humans,...
... and I am rooting for the machines.'
—Claude Shannon

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

2.1 ARTIFICIAL INTELLIGENCE

Definition 2.1. AI is the branch of Computer Science that studies general principles of intelligent agents and how to construct them [RND10].

[RND10] Russell et al., *Artificial Intelligence*.

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 intelligence was an intrinsic human characteristic.

In this dissertation:

Definition 2.2. Intelligence is the ability to predict a course of action to achieve success in specific goals.

[1] For a list with 70 definitions of intelligence, see [LH07].

2.1.2 *Intelligent Agents*

Under our generous definition, intelligence is not limited to humans. It applies to any agent[2]: animal or machine. For example, a bacteria can perceive its environment through chemical signals, process them, and then produce chemicals to signal other bacteria. An air-conditioning can observe temperature changes, 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 a calculator does not know arithmetics.

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

2.1.3 *A strange inversion of reasoning*

This competence without comprehension is what the philosopher Daniel Dennett calls *Turing's strange inversion of reasoning*[3]. The idea of a *strange inversion* comes from one of Darwin's 19th-century critics (MacKenzie as cited by Dennett):

[3] In his work, Turing discusses if computers can "think", meaning to examine if they can perform indistinguishably from the way thinkers do.

> *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

Counterintuitively to MacKenzie and many others to this date, intelligence can emerge from absolute ignorance. Turing's strange inversion of reasoning comes from the realisation 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 [Tur07].

[Tur07] Turing, 'Computing Machinery and Intelligence'.

2.2 DREAMING OF ROBOTS

2.2.1 *From mythology to Logic*

The idea of creating an intelligent agent is perhaps as old as humans. There are accounts of artificial intelligence in almost any ancient mythology: Greek, Etruscan, Egyptian, Hindu, Chinese [May18]. For example, in Greek mythology, the story of the bronze automaton of

[May18] Mayor, *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*.

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 particular, Aristotelian syllogism, *laws of thought*, provided patterns for argument structures to yield irrefutable conclusions, given correct premises. These ancient developments were the beginning of the field we now call *Logic*.

2.2.2 *Rationalism: The Cartesian view of Nature*

In the 13th century, the Catalan philosopher Ramon Lull wanted to produce all statements the human mind can think. For this task, he developed "logic paper machines", discs of paper filled with esoteric coloured diagrams that connected symbols representing statements. Unfortunately, according to Gardner, in a modern reassessment of his work, *"it is impossible, perhaps, to avoid a strong sense of anti-climax"* [Gar59]. With megalomaniac self-esteem that suggests psychosis, his delusional sense of importance is more characteristic of cult founders. On the bright side, his ideas and books exerted some magic appeal that helped them be rapidly disseminated through all Europe [Gar59].

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

Lull's work greatly influenced Leibniz and Descartes, who, in the 17th century, believed that all rational thought could be mechanised. This belief was the basis of **rationalism**, the epistemic view of the *Enlightenment* 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-evident*, and all truths can be derived from them.

There was considerable interest in developing artificial languages during this period. Nowadays, they are called formal languages.

> *If controversies were to arise, there would be no more need for 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 on society until today. In the 19th century, George Boole and others

[Gar59] Gardner, *Logic machines and diagrams*.

developed a precise notation for statements about all kinds of objects in Nature and their relations. Before them, Logic was philosophical rather than mathematical. The name of Boole's masterpiece, "*The Laws of Thought*", is an excellent indicator of his Cartesian worldview.

At 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 particular, Hilbert's *Entscheidungs Problem* (decision problem) asked if there were limits to mechanical Logic proofs [Cha06].

[Cha06] Chaitin, *Meta Math! The Quest for Omega*.

Kurt Gödel's incompleteness theorem (1931) proved that any language expressive enough to describe arithmetics of the natural numbers is either incomplete or inconsistent. This theorem imposes a limit on 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.

FIGURE 2.2: David Hume, Scottish Enlightenment philosopher, historian, economist, librarian and essayist.

Alan Turing brought a new perspective to the *Entscheidungs Problem*: a function on natural numbers that an algorithm in a formal language cannot represent cannot be computable [Cha06]. Gödel's limit appears in this context as functions that are not computable, *i.e.* no algorithm 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 Artificial Intelligence research as we know it. An area in which Turing himself was very much invested.

2.2.3 *Empiricism: The sceptical view of Nature*

The response to **rationalism** was **empiricism**, the epistemological view that knowledge comes from sensory experience, our perceptions of the world. Locke explains this with the peripatetic axiom[4]: "*there is nothing in the intellect that was not previously in the senses*" [Uzg20]. Bacon, Locke and Hume were great exponents of this movement, which established the grounds of the scientific method.

[4] This citation is the principle from the Peripatetic school of Greek philosophy and is found in Thomas Aquinas' work cited by Locke.

[Uzg20] Uzgalis, 'John Locke'. URL: https://plato.stanford.edu/archives/spr2020/entries/locke/

David Hume, in particular, presented in the 18th century a radical empiricist view: reason only does not lead to knowledge. In [Hum09], Hume distinguishes *relations of ideas*, propositions that derive from deduction and *matters of facts*, which rely on the connection of cause and effect through experience (induction). Hume's critiques, known

[Hum09] Hume, *Tratado da natureza humana*.

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

From Hume's own words:

> *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 nowise necessary.*
>
> — David Hume

There is no logic to deduce that the future will resemble the past. Still, we expect uniformity in Nature. As we see more examples of something happening, it is *wise* to expect that it will happen in the future just as it did in the past. There is, however, no *rationality*[5] in this expectation.

[5] In the philosophical sense.

Hume explains that we see conjunction repeatedly, *e.g.* "bread" and "nourish", and we expect *uniformity in Nature*; we hope that "nourish" will always follow "eating bread"; When we fulfil this expectancy, we misinterpret it as causation. In other words, we *project* causation into phenomena. Hume explained 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* [Hue17]: 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 beneficial and wise, but it is, *stricto sensu*, *irrational*.

[Hue17] Huebner, *The Philosophy of Daniel Dennett*.

The epistemology of the sceptical view of Nature is science: to weigh one's beliefs to the evidence. Knowledge is not absolute truth but justified belief. It is a Babylonian epistemology.

In rationalism, Logic connects knowledge and good actions. In empiricism, the connection between knowledge and justifiable actions is determined by probability. More specifically, Bayes' theorem. As Jaynes puts it, probability theory is the "Logic of Science" [Jay03]. [6]

[6] The Bayes' theorem is attributed to the Reverend Thomas Bayes after the posthumous 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 electronic units could be wired together, act and interact by physiologically plausible principles and perform complex logical calculations [RND10]. Moreover, they showed that any computable function could be computed by some network of connected neurons [MP43]. Their work marks the birth of Artificial Neural Networks (ANNs), even before the field of AI had this name. It was also the birth of **Connectionism**, using artificial neural networks, loosely inspired by biology, to explain mental phenomena and imitate intelligence.

FIGURE 2.3: Claude Shannon, father of "information theory".

Their work inspired John von Neumann's demonstration of how to create a universal Turing machine 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**, disregarding Connectionism.

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

[RND10] Russell et al., *Artificial Intelligence*.

[MP43] McCulloch and Pitts, 'A logical calculus of the ideas immanent in nervous activity'.

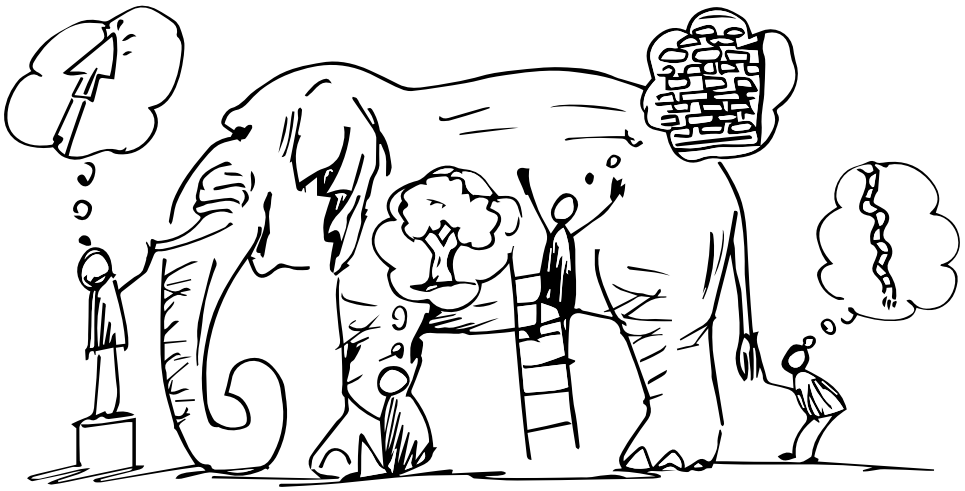


FIGURE 2.4: The Blind Men and the Elephant.

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 [Sax16]

2.3 BUILDING INTELLIGENT AGENTS

2.3.1 *Anatomy of intelligent agents*

Like the blind men in the parable, an intelligent agent shall model her understanding of Nature from limited sensory data.

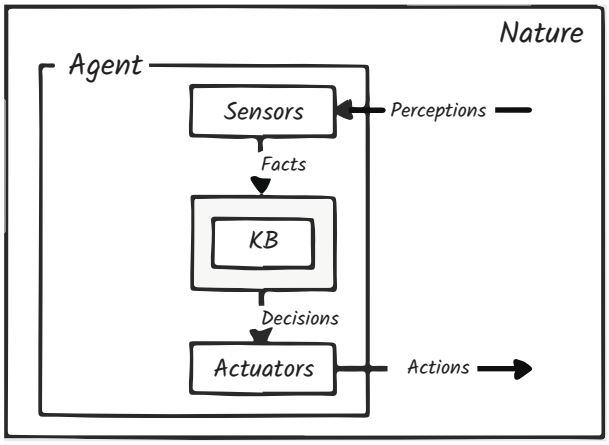


FIGURE 2.5: Anatomy of an Intelligent Agent. Inspired by art in [RND10].

Thus, an agent perceives her environment with sensors, treat sensory data as facts and use these facts to possibly update its model of Nature, use the model to decide her actions, and acts via her actuators. In a way, agents continually communicate with Nature 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 "thinks" the world is (her "mental picture" of the environment). Therefore, it should be consistent with it: if something is true in Nature, it is equally valid, *mutatis mutandis*, in the model. A Model should also be as simple as possible so that the agent can make decisions that maximise a chosen performance measure, but not simpler. As the agent knows more about Nature, less it gets surprised by it.

This rudimentary anatomy is flexible enough to entail different epistemic views, like the rationalist (mathematical) and the empiricist (scientific); different approaches to how to implement the knowledge base (it can be learned, therefore updatable, or it can be set in stone

from an expert prior knowledge); and also from how to implement it (a robot or 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 pinnacle 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, Symbolism dominated AI until the 1980s. It was so ubiquitous that symbolic AI is even called "good old fashioned AI" [RND10].

[RND10] Russell et al., *Artificial Intelligence*.

The symbolic approach can be traced back to Nichomachean Ethics [Ari00]:

[Ari00] Aristotle, *Aristotle: Nicomachean Ethics*.

> *We deliberate not about ends but 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 et al., 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 proof that there are other ways to build intelligent agents. Moreover, it is also a negation of competence without comprehension. It seems that even for AI researchers, the strange inversion of reasoning is uncomfortable (Chapter 1).

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 *ex post facto*? We humans tend to think our rational assessments lead to actions, but it is also likely possible that we act and then rationalise afterwards to justify what we have done, fullheartedly believing that the rationalisation came first.

Claude Shannon's Theseus

After writing what is probably the most important master's dissertation of the 20th century and "inventing" Information Theory, what made possible the Information Age we live in today, Claude Shannon enjoyed the freedom to pursue any interest to which his curious mind led him [SG17]. 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 most outstanding achievements in AI was Theseus, a robotic maze-solving mouse.[7]

[SG17] Soni and Goodman, *A mind at play: how Claude Shannon invented the information age*.

[7] Many AI students will recognise in Theseus the inspiration to Russel and Norvig's Wumpus World [RND10].

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.*

Symbolic AI problems

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

Anyhow, by 1965, there were already programs that could solve any solvable problem described in logical notation [RND10, p.4]. However, hubris and lack of philosophical perspective made computer scientists believe that "intelligence was a problem about to be solved[8]."

[RND10] Russell et al., *Artificial Intelligence*.

[8] Marvin Minsky, head of the artificial intelligence laboratory at MIT (1967)

Those inflated expectations lead to disillusionment and funding cuts [RND10]. 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: they did not count on the exponential explosion of their problems.

2.3.3 *Connectionism: a different approach*

The fundamental idea in Connectionism is that **intelligent behaviour emerges from a large number of simple computational units when networked together** [GBC16].

[GBC16] Goodfellow et al., *Deep Learning*.

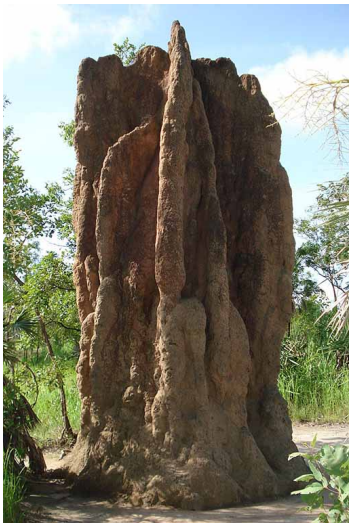
It was pioneered by McCulloch and Pitts in 1943 [MP43]. One of Connectionism's first wave developments was Frank Rosenblatt's Perceptron, an algorithm for learning binary classifiers, or more specifically threshold functions:

[MP43] McCulloch and Pitts, 'A logical calculus of the ideas immanent in nervous activity'.

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

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

(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.

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

From where does termites intelligence come?

> *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 do not think, but thinking arises in their connections.*
>
> — Radhika Nagpal, Harvard University [Mar16].

[Mar16] Margonelli, *Collective Mind in the Mound: How Do Termites Build Their Huge Structures?*.
URL: https://www.nationalgeographic.com/news/2014/8/140731-termites-mounds-insects-entomology-science/

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

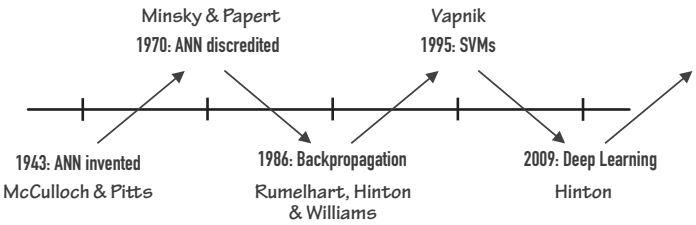


FIGURE 2.7: A brief history of connectionism. Adapted from [Tis20].

In contrast with the symbolic approach, in neural networks, the knowledge is not explicit in symbols but implicit in the strength 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.

FIGURE 2.8: Is this a cat?

2.3.4 *Machine Learning*

Look at Figure 2.8. Is this a picture of a cat? How to write a program to do such a simple classification task (cat/no cat)? One could develop clever ways to use *features* from the input picture and process them to guess. Though, it is not an easy program to design. Worse, even if one manages to program such a 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 (NLP), Speech Recognition; much mental effort was put, with inferior results, in problems that we humans solve with apparent ease.

The solution is an entirely different approach for building artificial intelligence: instead of making the program do the *task*, build the program that outputs the program that does the *task*. In other words,

learning algorithms use "training data" to infer the transformations to the input that generates the desired output.

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 [MRT12]:

[MRT12] Mohri et al., *Foundations of Machine Learning*.

- **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 unlabelled training data and makes predictions for unseen instances.
- **Semi-supervised learning:** The learner receives a training sample consisting of labelled and unlabelled data and makes predictions for unseen examples. Semi-supervised learning is usual in settings where unlabelled 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 2010s have been an AI Renaissance not only in academia but also in the industry. Such successes are mostly 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 to discover good representations, often at multiple levels, with higher-level learned features defined in terms of lower-level features."*
>
> *— Joshua Bengio [Ben12]*

[Ben12] Bengio, 'Deep learning of representations for unsupervised and transfer learning'.

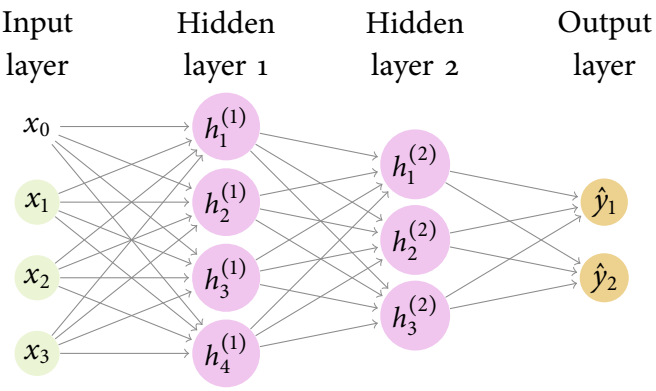
The name is explained by [GBC16]: "*A graph showing the concepts being built on top of each other is a deep graph. Therefore the name, deep learning*". Although it is a direct descendant of the connectionist

[GBC16] Goodfellow et al., *Deep Learning*.

movement, it goes beyond the neuroscientific perspective in its modern form. It is more a general principle of learning multiple levels of compositions.

The quintessential example of a deep learning model is the deep feedforward network or Multilayer Perceptron (MLP) [RND10].

[RND10] Russell et al., *Artificial Intelligence*.



Definition 2.3. Let,

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

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

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

σ be a nonlinear function,

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

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

a hidden layer is defined by:

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

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 (Section 4.3). Zhang et al. show state-of-the-art convolutional deep neural networks can easily fit a random labelling of training data [Zha+16].

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization.*

2.4 CONCLUDING REMARKS

This chapter derived the need for a *language* from the definitions of *intelligence* and *intelligent agents*. An intelligent agent needs *language* to store her knowledge (what she has learned) and with that to communicate/share this knowledge with its future self and with other agents.

We claim (without proving) that a language can be derived from a definition of knowledge: an epistemic choice. We claim that mathematics and science can be seen as languages that differ in consequence of different views on what knowledge is and gave historical background on two epistemic views, Rationalism and Empiricism (Sections 2.2.2 and 2.2.3).

We gave historical background on Artificial Intelligence (AI) and showed that different epistemic views relate to AI movements: Symbolism and Connectionism. We gave some background on basic AI concepts: intelligent agents, machine learning, types of learning, neural networks and deep learning, showing that DL relates to Connectionism and, hence, to science and an empiricist epistemology. Previously (Section 1.1.3), we have discussed that Computer Science generally relates to the rationalist epistemology. We hope this can help us better understand our research community.

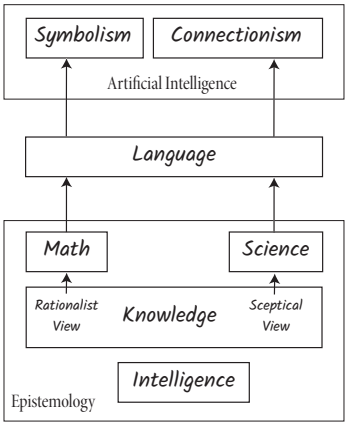


FIGURE 2.9: In this chapter we derived the need for a language from a philosophical axiom defining *intelligence*.

2.4.1 *Assumptions*

1. A definition of intelligence (Section 2.1.1)
2. An epistemic choice on the definition of Knowledge (Sections 2.2.2 and 2.2.3)

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 sceptical agents.

3.1 FROM LANGUAGE TO PROBABILITY

3.1.1 *Formal Languages*

We, as intelligent agents, do not know how Nature is; we only know how we perceive it. Our ideas are mental pictures of how we imagine Nature. Like in the story of the blind men and the elephant (Section 2.2.4), how do we know 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 Nature, our model, is consistent with the experience of others.[1]

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

We use language to describe Nature. 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 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 the chemical structures of molecules. Programming languages are formal languages that express computations. In a nutshell, a formal language is a set of words (strings) whose letters (symbols) are taken from an alphabet and are well-formed according to a specific set of rules, grammar. Let $L =< \Sigma, \Phi >$ be a formal language where:

$$\Sigma = \{S_1, S_2, \cdots, S_n\} \text{ is an alphabet,} \tag{3.1}$$

$$\Phi = \Phi_1 \cup \Phi_2 \cup \cdots \cup \Phi_k \text{ is a set of operations, the grammar,} \tag{3.2}$$

and:

$$\Phi_1 \text{ is the set of unary operations,}$$
$$\Phi_2 \text{ is the set of binary operations,}$$
$$\dots$$
$$\Phi_k \text{ is the set of k-ary operations.}$$

A formal language allows a quantitative description of a state of knowledge and defines how this state can be updated on new evidence.[2]

[2] An inference method defines the rules for updating knowledge.

With this definition, we can also think that a formal language is what Sowinski calls a *realm of discourse*, *i.e.* all the valid formed *strings*[3] that one can derive; everything one can *say* about Nature.

[3] Strings, words, sentences, propositions, formulae are names used interchangeably through the literature.

Interestingly, formal languages allow us to manipulate representations of the environment 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 at a superhuman speed and accuracy without ever comprehending what they are doing.

3.1.2 *From Rationalism to Propositional Calculus*

Rational Agents can form representations of a complex world, use deduction as the inference process 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 rational agent establishes a particular statement's truth value, all statements formed in her knowledge base from that statement instantly feel that update. Therefore, 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[4] can be either true or false;

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;

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

IV. **minimal**; uses the most reduced 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 [Sow16; Jay03]: NAND (or XOR), and it is also equivalent to Propositional Calculus.[5] In other words, Logic is the language that emerges from our desiderata, from rationalism. **Logic is the language of mathematics**.

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

[Sow16] Sowinski, 'Complexity and stability for epistemic agents: The foundations and phenomenology of configurational Entropy'.

[Jay03] Jaynes, *Probability Theory: The Logic of Science*.

[5] Proposition is synonym to sentence and Propositional Calculus is also known as Sentential Calculus.

3.1.3 *From Empiricism to Probability Theory*

The constraints that lead to Logic are very restrictive to use in the real-world; rational language has a comparatively small realm of discourse. Hume would say that it is only helpful 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 at least the empiricist perspective where knowledge is justified belief, and that one should *weigh 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., previous experience and evidence gathered.

Sceptical Agents In the sceptical agent, the one derived from the empiricist epistemology (authors have called these agents epistemic agents [Cato8], idealised epistemic agents [Sow16] or robots [Jay03]), beliefs are not independent of each other [Cato8], 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.

[Cato8] Caticha, *Lectures on Probability, Entropy, and Statistical Physics*.

Desiderata for a sceptical language As we did for rational agents, let us state a set of desired characteristics for the language of science, $L_S =< \Sigma_S, \Phi_S >$ [6]:

I. **Knowledge is a set of beliefs, quantifiable by real numbers and dependent on prior evidence [Sow16; Cato8; Jay03]:** 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. Thus we can list statements in an increasing

[6] [Sow16; Cato8; Jay03] also present this same idea of deriving probability theory from desiderata.

plausibility order. Real numbers can represent this transitive ordering.[7]

Let b be a measure of degrees of belief in S given some previous knowledge (or axiom) K:[8]

$$b : \Sigma_S \to \mathbb{R} \tag{3.3}$$
$$b : S \mapsto b(S|K) \tag{3.4}$$

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

II. **"Common sense:"**

The plausibility of compound sentences should be related by some logical function 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 there are such functions f and g that [Sow16]:

$$b(\neg S|K) = f[b(S|K)] \quad \text{(NOT)}$$
$$b(S_1 \wedge S_2|K) = g[b(S_1|K), b(S_1|S_2), b(S_2|K), b(S_2|S_1)] \quad \text{(AND)}$$

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 that are 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 (that will be seen in Section 3.4, Figure 3.1) [Sow16; Jay03; Cat08; TD15].

In other words, Probability Theory is the language that emerges from our desiderata, from empiricism. '**Probability theory is the**

[7] We are implicitly assuming that the language we are building has infinite statements. A further discussion on this continuity assumption can be found in [Sow16, p. 26].

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

[Sow16] Sowinski, 'Complexity and stability for epistemic agents: The foundations and phenomenology of configurational Entropy'.

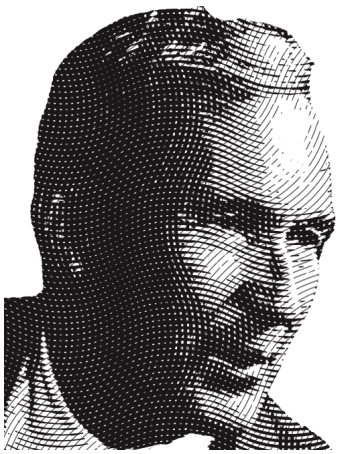


FIGURE 3.1: Andrey Kolmogorov, Soviet mathematician.

[Jay03] Jaynes, *Probability Theory: The Logic of Science*.

[Cat08] Caticha, *Lectures on Probability, Entropy, and Statistical Physics*.

[TD15] Terenin and Draper, 'Cox's Theorem and the Jaynesian Interpretation of Probability'.

Logic of Science' [Jay03], and our measure b is usually called probability P.

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

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[9] that humans are not rational, neither sceptical agents. The whole idea of imagining an epistemic agent is a consequence of addressing intelligence without human complexities.

[9] Unless you are an economist.

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 [Wie82]. Futhermore, if the human brain is not exposed to some samples in its infancy, it will never achieve the accuracy level 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 IIIa.

[Wie82] Wiesel, 'Postnatal Development of the Visual Cortex and the Influence of Environment'.

3.2 FORMALIZING PROBABILITY THEORY

We derived Cox's axioms from a list of desired properties of the language for sceptical agents. We also know that it is possible to derive Kolmogorov's Axioms (which will be defined soon in Section 3.4) from those axioms. In the next sections, we will use the Kolmogorov Axioms to formalise Probability theory.

Several concepts in the following sections are *relations of ideas*, not *matters of fact*. For example, the probability of an *event* E, P(E), can be computed by marginalisation (as we will show in Section 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 now that we will expose the idealised development of 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 **realisation** is a point $\omega \in \Omega$:

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

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

$$A = \{\boldsymbol{a}_1 = \boxdot, \boldsymbol{a}_2 = \boxdot, \boldsymbol{a}_3 = \boxdot\}$$
$$A \subset \Omega$$

We say that $A_1, A_2, \cdots$ are **mutually exclusive** or **disjoint** *events* if $A_i \cap A_j = \varnothing, \forall\, i \neq j$. 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).

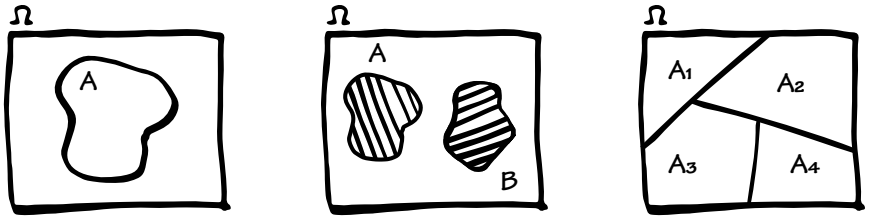


FIGURE 3.2: Events, disjoint events and partitions.

(a) An *event* A.

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

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

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

$$A_1, A_2, \cdots A_i \text{ s.t. } (A_1 \cup A_2 \cup A_3 \cdots = \bigcup_{i=1}^{\infty} A_i) = \Omega \tag{3.5}$$

3.4 KOLMOGOROV'S DEFINITION OF PROBABILITY

Definition 3.1 (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 Kolmogorov's axioms [Was13]:

[Was13] Wasserman, *All of statistics: a concise course in statistical inference.*

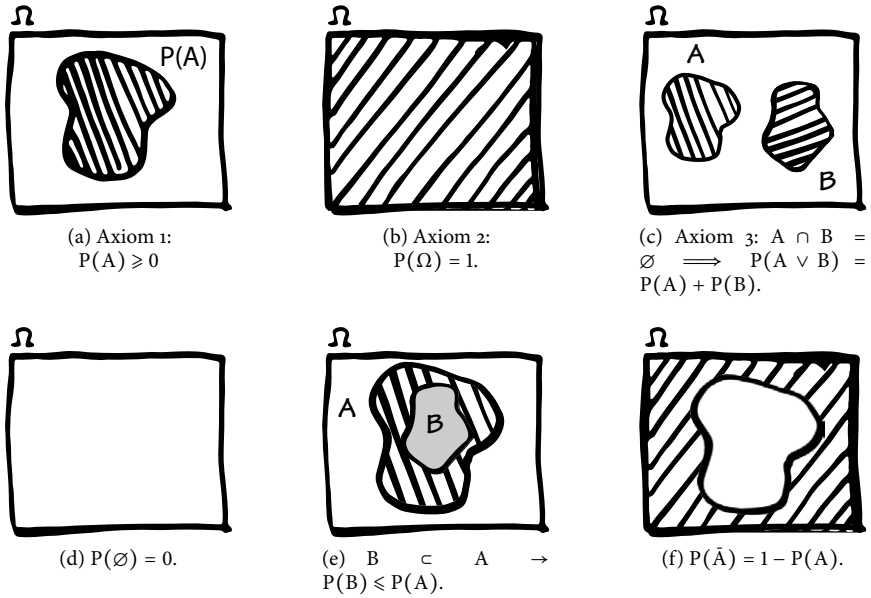
Axiom 1. $P(A) \geqslant 0, \forall A$

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

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

$$P(A \vee B) = P(A) + P(B) \qquad \text{(Sum Rule)}$$

Visually, we can represent the probability of an *event* A, $P(A)$, as the proportion of the sample space the *event* occupies. To differentiate *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: $A \cap B = \varnothing \implies P(A \vee B) = P(A) + P(B)$.

(d) $P(\varnothing) = 0$.

(e) $B \subset A \rightarrow P(B) \leqslant P(A)$.

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

FIGURE 3.3: Kolmogorov's Axioms and their direct consequences.

Directly from the Kolmogorov Axioms, one can derive [Jay03] other properties (see Figures 3.3a to 3.3c):

$$P(\varnothing) = 0 \tag{3.6}$$

$$B \subset A \implies P(B) \leqslant P(A) \tag{3.7}$$

$$0 \leqslant P(A) \leqslant 1 \tag{3.8}$$

$$P(\bar{A}) = 1 - P(A). \tag{3.9}$$

[Jay03] Jaynes, *Probability Theory: The Logic of Science*.

3.5 JOINT EVENT

Definition 3.2. A joint *event* A, B is the set of outcomes where:

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

Therefore,

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

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 $P(A, B)$.

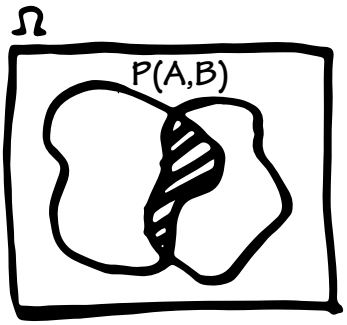


FIGURE 3.4: Joint event (A, B).

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

3.6 INDEPENDENT EVENTS

Definition 3.3. Events A and B are **independent** ($A \perp B$) if:

$$A \neq \varnothing, B \neq \varnothing \implies P(A) > 0, P(B) > 0 \tag{3.10}$$

$$P(A, B) = P(A \wedge B) = P(A) \cdot P(B) \tag{3.11}$$

(Product Rule)

Disjoint *events* cannot be independent, since (from (3.10)) $P(A) \cdot P(B) > 0$, but as disjoint *events* (Figure 3.2b) $P(A \wedge B) = P(\varnothing) = 0$, leading to contradiction.

Independence can be assumed or derived by verifying:

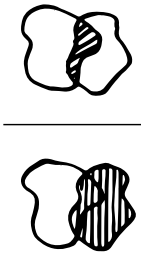
$$P(A \wedge B) = P(A) \cdot P(B). \tag{3.12}$$

(Independent variables)

3.7 CONDITIONAL PROBABILITY

As we have explained before (Section 3.1.3), the plausibility of an outcome or a set of outcomes depends on a web of interconnected prior beliefs. So, what exists are probabilities *conditional* to a given prior assumption.

$$P(A|B) = \frac{\text{}}{\text{}}$$



Definition 3.4. If $P(B) > 0$ then the **conditional probability** of A given B is:

$$P(A|B) \triangleq \frac{P(A, B)}{P(B)} \tag{3.13}$$

$$P(A, B) \triangleq P(A|B) \cdot P(B) \tag{3.14}$$

Except if $P(A) \equiv P(B), P(A|B) \neq P(B|A)$. Also, $P(A|B) = P(A) \iff A \perp B$.[10]

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

3.8 MARGINAL PROBABILITY

Theorem 3.1. *Let* $A_1, \cdots, A_k$ *be a partition of* Ω. *Then, for any* event *B,*

$$P(B) = \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i) \tag{3.15}$$

Proof. [11] Define $C_i = (B, A_i)$. Let $C_1, \cdots C_k$ be disjoint and $B = \bigcup_{i=1}^{k} C_i$. Therefore:

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

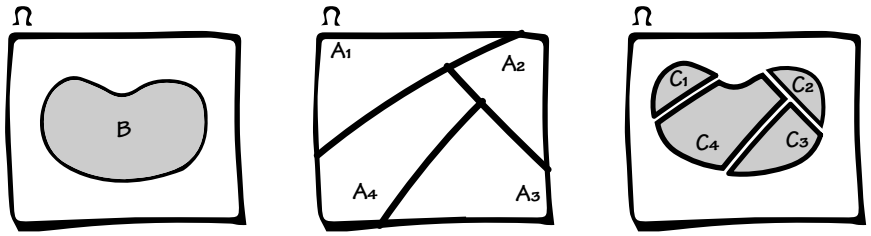


FIGURE 3.5: An *event* B, a partition A_i over Ω, and $C_i = (B, A_i)$.

$$P(B) \triangleq P(\bigcup_{i=1}^{k} C_i) \overset{\text{Sum Rule}}{=} \sum_i P(C_i) \tag{3.16}$$

$$\triangleq \sum_i P(B, A_i) \overset{3.13}{=} \sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)$$

(Law of Total Probability)

□

3.9 BAYES' THEOREM

Theorem 3.2 (Bayes' theorem)**.** *Let* $A_1, \cdots, A_k$ *be a partition of* Ω *s.t.* $P(A_i) > 0, \forall i$ *then,* $\forall i = 1, \cdots, k$*:*

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

Proof. From equations 3.13, 3.14 and 3.15:

$$P(A_i|B) \overset{3.13}{=} \frac{P(A_i, B)}{P(B)} \overset{3.14}{=} \frac{P(B|A_i) \cdot P(A_i)}{P(B)} \tag{3.18}$$

$$\overset{3.15}{=} \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} \tag{3.19}$$

□

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

3.10 RANDOM VARIABLES

Definition 3.5. A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω, $\omega \mapsto X(\omega)$.

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

$$P(X = x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) = x\}) \tag{3.20}$$

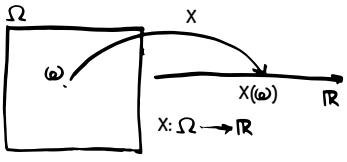


FIGURE 3.6: Random variable.

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.

The difference between a random variable X and its "realisation" is the difference between a distribution and a sample from that distribution. In particular, a random variable X is "formalised" in terms of a function from the sample space to some result space, typically $\mathbb{R}$. The realisation of a random variable is "what you get" when an *experiment* is run, and you apply X to *events* that happened.

3.10.1 *Notation abuse*

If a *random variable* is a function, how can we write $\mathrm{P}(X = 4)$ or $\mathrm{P}(X > 7)$? Such confusion is due to some notation abuse that became standard in works on probability theory. It is not easy to grasp it initially, but the explanation was already stated at (3.20). $\mathrm{P}(X = x)$ is a shorthand for $\mathrm{P}(X^{-1}(x))$.

Technically, a *random variable* is a function. In practice, it is just a mathematical tool to help us associate propositions with numbers. It is called a *random variable* because the notation abuse treats 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 proposition "*The dice will roll to number 4.*" plausibility 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 $\mathrm{P}(\mathrm{D}) = \frac{1}{6}$. Now, we are going one step further; instead of using the *event* D we use the *random variable* D, in italic, and say $\mathrm{P}(D = 4) = \frac{1}{6}$.

Notice the difference between a *random variable* and an *event*:[12] D could assume any value (even $D = 7$, which is 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.

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, which we cannot cope with *events*.

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

$$D = \mathbb{1}_{\mathrm{D}}$$

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

3.11 PROBABILITY DISTRIBUTIONS

Definition 3.6. 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):

$$p(x) = \mathrm{P}(X = x), \tag{pmf}$$

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

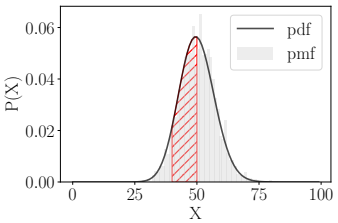


FIGURE 3.7: Probability mass function, probability density function, and probability of an interval (hatched area).

Definition 3.7. 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 $\mathrm{A} = [a, b]$, as the area under its curve for the interval A:

$$\mathrm{P}(\mathrm{A}) = \mathrm{P}[a \leqslant X \leqslant b] = \int_a^b p(x)\,dx, \text{ and:} \tag{3.21}$$

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

Now that we explained what distributions are,[13] here we highlight some useful distributions:

[13] In this dissertation, we will use $\mathrm{P}(X)$ to express the probability of a random variable, and $p(x)$ to represent a *pmf* or *pdf* of the random variable outcomes.

3.11.1 *Statistical model*

A statistical model is a function $p_\theta(x) \equiv p(x|\theta)$ representing the relationship between a parameter[14] θ and potential outcomes x of a random variable X. In practice, we usually define a statistical model of a stochastic process for which we do not know the real distribution. Therefore, the parameter θ has to be inferred from the observed data.

[14] In this dissertation we are interested in vector-valued θ.

3.11.2 *Uniform distribution*

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

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

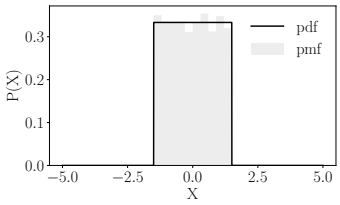


FIGURE 3.8: Uniform distribution.

3.11.3 *Normal distribution*

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

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

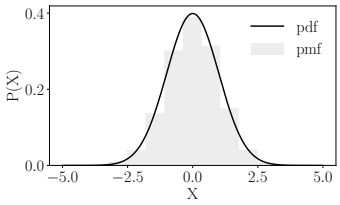


FIGURE 3.9: Gaussian distribution, also known as the *normal*.

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

3.11.4 *Exponential distribution*

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

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

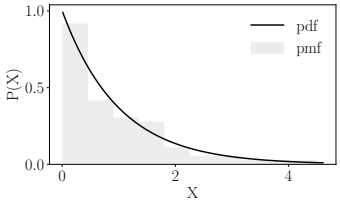


FIGURE 3.10: Exponential distribution.

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

3.12 JOINT DISTRIBUTIONS

Definition 3.8. Given a pair of discrete random variables X and Y, we define the **joint mass function** by $p(x, y) = \text{P}(X = x, Y = y)$.

Definition 3.9. 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. $\iint_{-\infty}^{\infty} p(x, y)\, dxdy = 1$

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

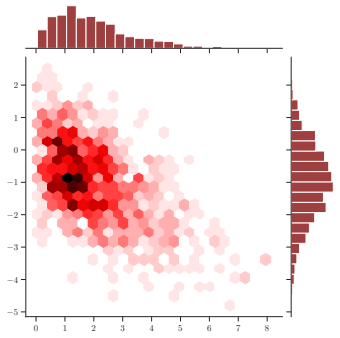


FIGURE 3.11: A chart of a joint distribution.

3.13 EXPECTANCY, VARIANCE AND COVARIANCE

Definition 3.10. The **expected value** or **mean** of X is:

$$\mathbb{E}(X) = \langle X \rangle = \sum_x x\, p(x)\, dx = \mu = \mu_X \tag{3.22}$$

Theorem 3.3. *Let $X_1, \cdots, X_n$ be random variables and $a_1, \cdots, a_n$ be constants, then from the* Sum Rule*:*

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i(\mathbb{E}(X_i)) \tag{3.23}$$

Theorem 3.4. *Let $X_1, \cdots, X_n$ be independent random variables, then from the* Product Rule*:*

$$\mathbb{E}(\prod_i X_i) = \prod_i \mathbb{E}(X_i) \tag{3.24}$$

Definition 3.11. Let X be a random variable with mean μ. The **variance** of X is defined by:

$$\sigma^2 = \sigma_X^2 = \mathbb{E}(X - \mu)^2 \tag{3.25}$$

assumming this expectation exists. The standard deviation is σ.

Definition 3.12. 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 [Was13, p.74]:

[Was13] Wasserman, *All of statistics: a concise course in statistical inference.*

$$\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y)) \tag{3.26}$$

and the correlation as:

$$\rho = \rho_{X,Y} = \rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{3.27}$$

Theorem 3.5. *The covariance satisfies:*

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{3.28}$$

3.14 INDEPENDENT SAMPLING

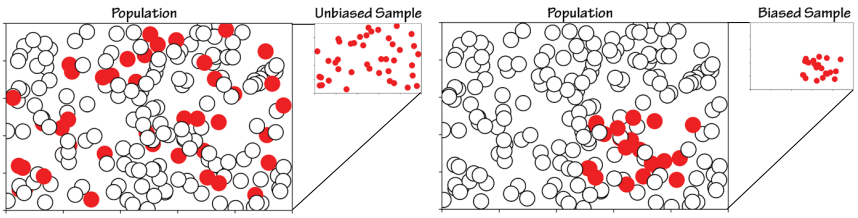


FIGURE 3.12: An i.i.d. sample (left) and a biased sample (right). Adapted from [MP18].

A *sample* is a set of examples[15] 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 drawn in the following observation.

[15] In this dissertation, an element of a sampling is called an *example*.

The i.i.d. assumption is useful wherever a census of the population of interest, knowing all possible values, is unfeasible. In this 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 3.12 — left).

If there is a constraint on which examples of the population are sampled, we say that the sample is *biased* (Figure 3.12 — right).

3.15 CONCLUDING REMARKS

This chapter derived *Logic* from the definition of knowledge as absolute truth and *Probability Theory* from knowledge as justified beliefs (Sections 3.1.2 and 3.1.3). To remind that our definition of knowledge is the basis for the Bayesian perspective of probability and that inference methods are languages, we can say (and prefer) that we derived *Bayesian inference* as the language of science. We proved what we claimed in the previous chapter (Chapter 2).

We needed to define *formal languages* (Section 3.1.1) and assume desiderata for the languages we wanted to build formally (Sections 3.1.2 and 3.1.3). We called *rational agents* the epistemic agents that use Logic as its inference method, and *sceptical agents* use Bayesian inference.

We found out that the desiderata for the sceptical language are equivalent to Cox's axioms (Section 3.1.3). From Cox's axioms, it is possible to derive Kolmogorov's axioms of Probability Theory. Which made us conclude that Bayesian inference is the language of science.[16]

From the derivation, we did a basic Statistics review (influenced by [Was13]). Many essential topics were left out from this short review chapter, where the focus was to present the concepts that we will use later on in this dissertation.

[16] Our definition of knowledge hinted at a Bayesian perspective of knowledge.

[Was13] Wasserman, *All of statistics: a concise course in statistical inference*.

3.15.1 *Assumptions*

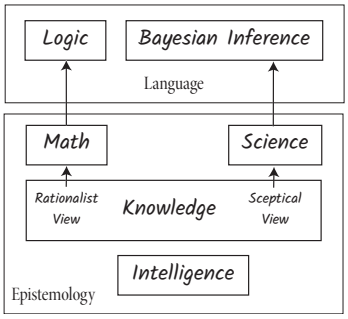


FIGURE 3.13: This chapter derives *Logic* as the language of Mathematics, in which knowledge is absolute; and *Bayesian inference* as the language of Science, in which knowledge is justified belief.

1. A definition of intelligence (Section 2.1.1);
2. A epistemic choice on the definition of Knowledge (Sections 2.2.2 and 2.2.3);
3. A definition of formal language;
4. Common assumptions of the epistemic agent language:
    a) consistency (Section 3.1.3, Item III and Section 3.1.2, Item III);
    b) minimality (Section 3.1.2, Item IV).
5. Assumption of the rational agent language:
    a) knowledge is absolute, a set of true or false sentences (Section 3.1.2, Item I);
    b) the language must be unambiguous (Section 3.1.2, Item II).
6. Assumption of the sceptical agent language:

a) Knowledge is a set of beliefs, quantifiable by real numbers and dependent on prior evidence (Section 3.1.3, Item I);

b) Common sense: The plausibility of compound sentences should be related by some logical function to the plausibility of the sentences that form them (Section 3.1.3, Item II).

As we have settled that our focus is Deep Learning, which relates to the sceptical agent, we will abstain from keeping the rational language assumptions in our analysis and assume an epistemic agent is sceptical.

4

Machine Learning Theory

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

—Andrey Kolmogorov

In which 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.

This chapter is influenced by the online lecture *Statistical Learning Theory - a Hitchhiker's Guide (NeurIPS 2018)* [STR18], the online lecture series *Statistical Learning Theory* [Mel18] and the book *Machine learning: a practical approach on the statistical learning theory* [MP18].

[STR18] Shawe-Taylor and Rivasplata, *Statistical Learning Theory - a Hitchhiker's Guide (NeurIPS 2018)*.
URL: https://youtu.be/m8PLzDmW-TY

[Mel18] Mello, *Statistical Learning Theory*.
URL: https://youtu.be/KTrRap4Spd0

[MP18] Mello and Ponti, *Machine learning: a practical approach on the statistical learning theory*.

4.1 MOTIVATION

As already discussed, learning is inferring general rules to perform a specific task by observing limited examples. Therefore, the learning algorithm must perform well in the sample already seen and, 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 ensure 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 1960s, with the work of Vapnik and Chervonenkis (see Figure 4.1). 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). We will also limit our overview of MLT to the

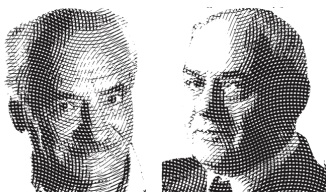


FIGURE 4.1: Chervonenkis (Left) and Vapnik (Right).

context of supervised binary classification problems. This limitation is not a deficiency of the theory but a mere choice of scope for this document.

4.2 THE LEARNING PROBLEM

The goal of learning is to understand Nature from experience, coming up with a theory, a tested hypothesis. A *concept* c is an 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). The convention is that labels are binary, $\mathcal{Y} = \{-, +\}$, therefore, we can assign the true label to the presence of the element, $c \subset \mathcal{X}$.

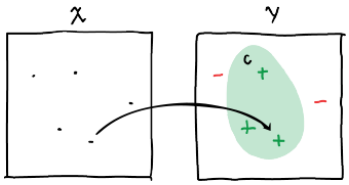


FIGURE 4.2: A *concept* c is an idealised input to output mapping, $\mathcal{X} \mapsto \mathcal{Y}$.

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

4.2.1 *The learning problem setting*

Supervised learning has three main components (see Figure 4.3):

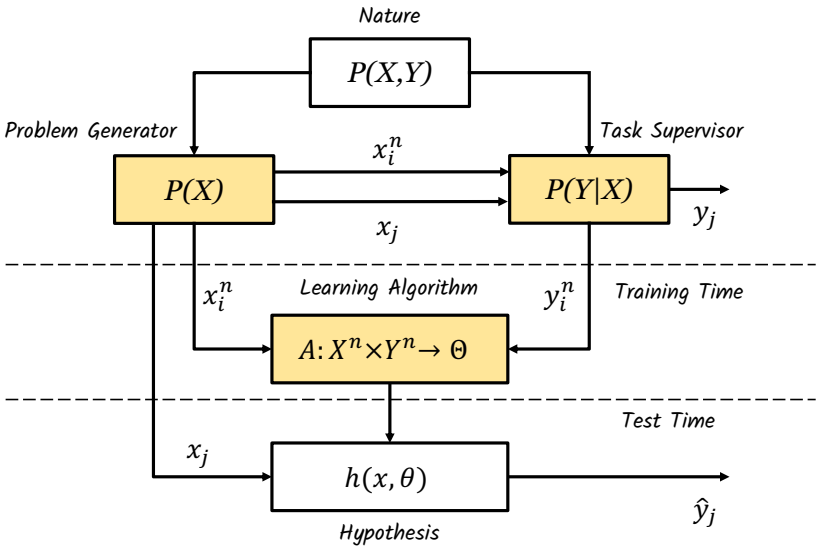


FIGURE 4.3: Learning problem setting.

1. A **generator** of vectors randomly draw from a probability distribution $\mathrm{P}(X)$, $x \sim \mathrm{P}(X), x \in \mathcal{X}$,[1] which represent instances of the problem[2];

[1] In Chapter 5, we use $\mathbb{A}_X$ to represent the domain of X to emphasise that the domain is finite; it is an alphabet. Here we use $\mathcal{X}$ to remember that this domain possibly is infinite.

[2] $\mathrm{P}(X = x_i) = \mathrm{P}_X(x_i) = \sum_j \mathrm{P}_{XY}(x_i, y_j) \therefore \mathrm{P}_{\mathcal{X}}$ is just a consequence of $\mathrm{P}(X, Y) \therefore x \sim \mathrm{P}_{\mathcal{X}} \equiv x \sim \mathrm{P}_{XY}$.

2. A **task supervisor** that knows the concept and returns an output vector y_i for every input vector x_i:

$$y_i = c(x_i), y_i \sim \mathrm{P}(Y \mid X = x_i). \tag{4.1}$$

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

$$\mathcal{A} : \underbrace{(\mathcal{X} \times \mathcal{Y})^n}_{S^n} \to \mathcal{H}. \tag{4.2}$$

[3] Hypothesis spaces will be explained in Section 4.2.3.

The problem of learning is choosing from the *hypothesis space* 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 $\mathrm{D} = \mathrm{P}(X, Y)$.

4.2.2 *Assumptions*

The common assumptions are as follows [MP18; VLS11]:

[MP18] Mello and Ponti, *Machine learning: a practical approach on the statistical learning theory*.

[VLS11] Von Luxburg and Schölkopf, 'Statistical learning theory: Models, concepts, and results'.

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

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

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

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

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

4.2.3 *Hypothesis spaces*

The problem setting relies on the idea of a *hypothesis space* (also known as a *hypothesis class*). A hypothesis space is the set of all hypotheses[4] a functional learning algorithm $\mathcal{A}$ can generate. In the same hypothesis space $\mathcal{H}$, hypotheses differ by their parameter vector θ. Choosing a hypothesis h_i is choosing its parameter θ_i.

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

$$h : \mathcal{X} \times \Theta \to \mathcal{Y}, \tag{4.3}$$

$$h(x) = p(y \mid x \wedge \theta), \ \theta \in \Theta. \tag{4.4}$$

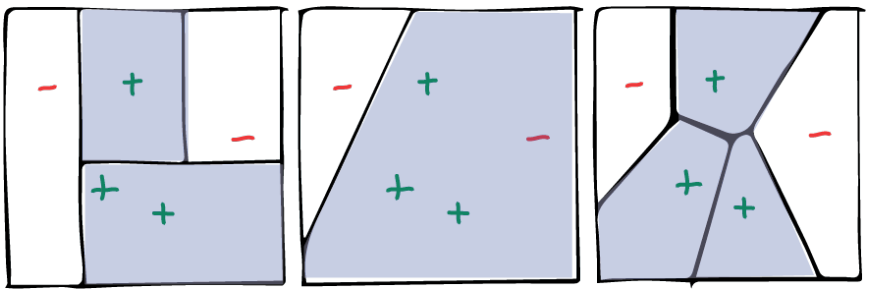


FIGURE 4.4: Different hypothesis spaces solutions for the same sample.

Different learners will constraint the input space $\mathcal{X}$ differently (see Figure 4.4). Some algorithms are more complex than others, meaning they can express more different functions.[5]

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

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

4.2.4 *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.

$$\mathrm{R}_{\mathrm{D}}(\theta) = \mathbb{E}(\ell(x, y, h(x, \theta)), (x, y) \sim \mathrm{D}, \theta \in \Theta \tag{4.5}$$

$$\theta_* = \arg\min_{\theta \in \Theta} \mathrm{R}(\theta) \tag{4.6}$$

$$h(x, \theta_*) = h_{\text{Bayes}} = \arg\min_{h \in \mathcal{H}_{\text{all}}} \mathrm{R}(h) \tag{4.7}$$

The risk R_{D} is also called the absolute (or out-of-sample or theoretical) error of the hypothesis.[6] Nevertheless, there is one crucial caveat: **the choice of the loss metric is arbitrary, which curbs any objective, metric independent, interpretation of the results.**

[6] R, $\mathrm{R}(\theta)$ and $\mathrm{R}(h)$ are used interchangeably in this dissertation.

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:

$$\hat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n}(\ell(x_i, y_i, h(x_i)), (x, y) \sim S \tag{4.8}$$

With this empirical risk $\hat{R}_S$ that we can evaluate, we find the hypothesis that minimises it. Given a sample $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, a hypothesis space $\mathcal{H}$, and a loss function ℓ, we define $h_\mathcal{H}$ as the function:

$$h_\mathcal{H} = \arg\min_{h \in \mathcal{H}} \hat{R}_S(h) \tag{4.9}$$

According to the law of large numbers (Section 4.5.3), if the sample is large enough, by induction, a hypothesis generated optimising $\hat{R}_S$ is close to R. However, it is essential to notice that we still have to discuss at which rate does $\hat{R}_S$ converge to R with regards to the the sample size.

4.3 BIAS-VARIANCE TRADE-OFF

When we define a subset of $\mathcal{H} \subset \mathcal{H}_{all}$ where to look for our hypothesis, we impose 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 $\mathcal{H}_{NN} \supset \mathcal{H}_{Perceptron}$.

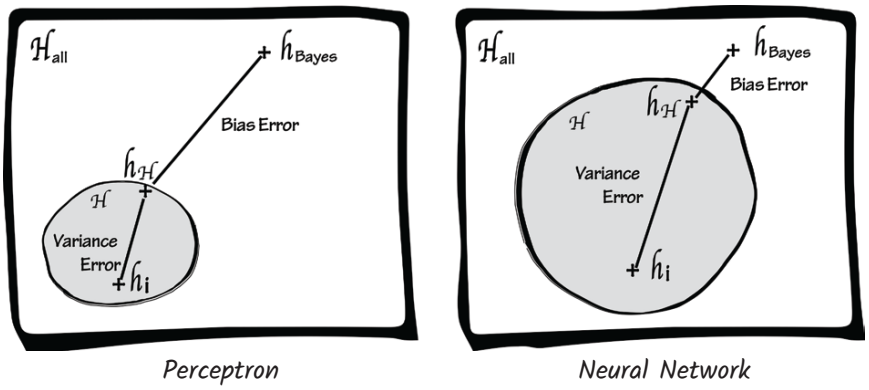


Figure 4.5: Bias and variance errors.

Accordingly, we can distinguish two 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 4.5 Perceptron). 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 (see Figure 4.5 Neural Network).

- **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 4.5 Neural Network).

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

$$\Delta(h_i) = \mathrm{R}(h_i) - \mathrm{R}(h_{\text{Bayes}}) \tag{4.10}$$

$$= \underbrace{(\mathrm{R}(h_i) - \mathrm{R}(h_{\mathcal{H}}))}_{\text{Variance Error}} + \underbrace{(\mathrm{R}(h_{\mathcal{H}}) - \mathrm{R}(h_{\text{Bayes}}))}_{\text{Bias Error}} \tag{4.11}$$

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

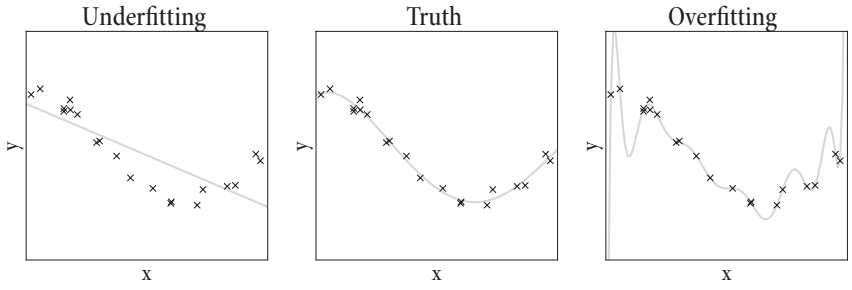


FIGURE 4.6: Example of underfitting and overfitting in a regression problem.

- **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 significant, 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 more substancial the bias, the smaller is the $\mathcal{H} \subset \mathcal{H}_{\text{all}}$, smaller is the variance error, but the more significant is the bias error; and vice-versa

Figure 4.7: Generalisation gap.

(Figure 4.7). This trade-off is the central paradigm of Machine Learning Theory [Sloo2], its crucial challenge, and has different names underfitting-overfitting, precision-complexity, and performanceprediction. **The goal of machine learning algorithms is to come up with the simplest model that explains the data, but not simpler**.

[Sloo2] Slonim, 'The information bottleneck: Theory and applications'.

> *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.*
>
> — *Avrim Blum [Bluo7]*

4.4 THE PAC LEARNING MODEL

Up to this point in the chapter, we have described MLT following 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 4.8) in 1984 [Val84]. The lack of citation to Vapnik and Chervonenkis literature is an indication that the overlap of CoLT and STL was reinvented. As expected, CoLT looks at the learning problem from a computational perspective, while SLT from a statistical one.

Figure 4.8: Leslie Valiant received the Turing Award in 2010.

"The PAC framework deals with the question of learnability for a concept class $\mathbb{C}$ and not a particular concept" [MRT12], where a concept class is a set of concepts c_i. The PAC model classifies concept classes in terms of their complexities to achieve an approximate solution; sample complexity, the number of examples needed, computation complexity, the number of iterations needed.

[Val84] Valiant, 'A theory of the learnable'.

[MRT12] Mohri et al., *Foundations of Machine Learning*.

In the PAC framework, a concept class $\mathbb{C}$ 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.

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

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

$$\exists h \in \mathcal{H} : \Pr_{x \sim D}[c(x) \neq h(x)] = 0 \rightarrow \text{learning}. \tag{4.12}$$

Nevertheless, this definition is too restrictive. For instance, if $c \notin \mathcal{H}$, there is no way for any h to perfectly imitate c. So let us redefine learning with new relaxed constraints to the absolute error:

$$\Pr_{x\sim D}[c(x) \neq h(x)] = R_D(h) \tag{4.13}$$

$$\exists h \in \mathcal{H} : \; R_D(h) \leqslant \epsilon, \; 0 < \epsilon < \tfrac{1}{2} \rightarrow \text{learning}. \tag{4.14}$$

Allowing some tolerance to error, however, is still not sufficient. On one side, a *hypothesis* does not need to be equal to the *concept* to be ***consistent* to the sample**, *i.e.* to correctly predict every example of the sample. In the figure Figure 4.9, 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 $\mathcal{X}$.

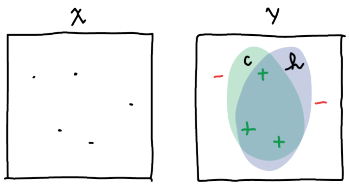


FIGURE 4.9: The concept versus the hypothesis.

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

$$S_n = \{(x_1, y_1), \cdots, (x_n, y_n)\} \sim D^n \tag{4.15}$$

is *unlucky*, and is a set of *bad* 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:

$$\exists h \in \mathcal{H}, \; 0 < \epsilon < \tfrac{1}{2}, \; 0 < \delta < \tfrac{1}{2} :$$
$$\Pr_{S\sim D^n}[R_D(h) > \epsilon] < \delta \rightarrow \text{learning}. \tag{4.16}$$

Nevertheless, if achieving such thresholds demands an unreasonable amount of data and time, can we 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 size$(c)$ the computational cost of representing a concept $c \in \mathbb{C}$.

Definition 4.1. A concept class $\mathbb{C}$ is **PAC-learnable** if there is a learning algorithm $\mathcal{A}$ and a polynomial function 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 $n \geqslant$ poly$(\frac{1}{\epsilon}, \frac{1}{\delta}, d, \text{size}(c))$ [MRT12]:

$$\Pr_{S\sim D^n}[R_D(h) \leqslant \epsilon] \geqslant (1 - \delta). \tag{4.17}$$

If $\mathcal{A}$ further runs in 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 $\mathbb{C}$ [MRT12].

[MRT12] Mohri et al., *Foundations of Machine Learning*.

4.5 PAC BOUNDS

As we stated before, one of the main goals of 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 of how this theoretical development allows us to make claims on unknown distributions and unseen examples.

4.5.1 *Guarantees for finite hypothesis spaces — consistent case*

Theorem 4.1 ([Hau88], finite space, consistent case)**.** *Let $\mathcal{H}$ be a finite hypothesis space, $\mathcal{A}$ a learning algorithm that returns a consistent hypothesis h, i.e. $\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0$, for any hypothesis $h \in \mathcal{H}$ and unknown distribution $\mathrm{D} = \mathrm{P}(X, Y)$.*
Let $|\mathrm{S}| = n$, then, $\forall n \geqslant 1$:

$$Pr\left[\exists h \in \mathcal{H} : \mathrm{R}_{\mathrm{D}}(h) > \epsilon\right] \leqslant |\mathcal{H}| e^{-\epsilon n} \tag{4.18}$$

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

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

$$\Pr_{x_j \sim \mathrm{S}}\left[\left(c(x_j) \neq h_{\text{bad}}(x_j)\right) = \varnothing\right] \leqslant (1 - \epsilon) \tag{4.19}$$

$$\Pr_{x_j \sim \mathrm{S}}\left[\mathrm{R}_{x_j}(h_{\text{bad}}) = 0\right] \leqslant (1 - \epsilon) \tag{4.20}$$

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

$$\Pr_{x_1 \sim \mathrm{S}}\left[\mathrm{R}_{x_1}(h_{\text{bad}}) = 0\right] \wedge \tag{4.21}$$

$$\Pr_{x_2 \sim \mathrm{S}}\left[\mathrm{R}_{x_2}(h_{\text{bad}}) = 0\right] \wedge \tag{4.22}$$

$$\cdots$$

$$\Pr_{x_n \sim \mathrm{S}}\left[\mathrm{R}_{x_n}(h_{\text{bad}}) = 0\right] \leqslant \underbrace{(1 - \epsilon) \cdots (1 - \epsilon)}_{n} \tag{4.23}$$

$$\Pr\left[\left(\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0\right) \wedge \left(\mathrm{R}_{\mathrm{D}}(h) > \epsilon\right)\right] \leqslant (1 - \epsilon)^n \tag{4.24}$$

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

$$\Pr\left[h_1 \in \mathcal{H}_{\text{bad}} : \left(\hat{\mathrm{R}}_{\mathrm{S}}(h_1) = 0\right) \wedge \left(\mathrm{R}_{\mathrm{D}}(h) > \epsilon\right)\right] \vee \tag{4.25}$$

$$\Pr\left[h_2 \in \mathcal{H}_{\text{bad}} : \left(\hat{\mathrm{R}}_{\mathrm{S}}(h_2) = 0\right) \wedge \left(\mathrm{R}_{\mathrm{D}}(h) > \epsilon\right)\right] \vee \tag{4.26}$$

$$\cdots$$

$$\vee \Pr\left[h_k \in \mathcal{H}_{\text{bad}} : \left(\hat{\mathrm{R}}_{\mathrm{S}}(h_k) = 0\right) \wedge \left(\mathrm{R}_{\mathrm{D}}(h) > \epsilon\right)\right] \leqslant \sum_{1}^{k} (1 - \epsilon)^n \tag{4.27}$$

$$\Pr\left[\exists h \in \mathrm{H} : \left(\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0\right) \wedge \left(\mathrm{R}_{\mathrm{D}}(h) > \epsilon\right)\right] \leqslant k(1 - \epsilon)^n \tag{4.28}$$

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

$$\begin{aligned} \Pr\left[\exists h \in \mathcal{H} : \mathrm{R}_{\mathrm{D}}(h) > \epsilon\right] &\leqslant |\mathcal{H}|(1-\epsilon)^n \\ (1-x) \leqslant e^{-x}, 0 \leqslant x \leqslant 1 &\implies \\ \Pr\left[\exists h \in \mathcal{H} : \mathrm{R}_{\mathrm{D}}(h) > \epsilon\right] &\leqslant |\mathcal{H}| e^{-\epsilon n} \end{aligned} \tag{4.29}$$

□

From the PAC framework:

$$\Pr\left[\exists h \in \mathcal{H} : \mathrm{R}_{\mathrm{D}}(h) > \epsilon\right] < \delta \tag{4.30}$$

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

$$\delta > |\mathcal{H}| e^{-\epsilon n} \geqslant \Pr\left[\exists h \in \mathcal{H} : \mathrm{R}_{\mathrm{D}}(h) > \epsilon\right] \tag{4.31}$$

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

Theorem 4.2 ([Hau88], finite space, consistent case: sample complexity)**.** *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.

$$\delta > |\mathrm{H}| e^{-\epsilon n} \qquad \text{(from (4.31))}$$

$$e^{-\epsilon n} < \frac{\delta}{|\mathcal{H}|} \tag{4.32}$$

$$-\epsilon n < (\ln\delta - \ln|\mathcal{H}|) \tag{4.33}$$

$$\epsilon n < (\ln|\mathcal{H}| - \ln\delta) \tag{4.34}$$

$$n < \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta}) \tag{4.35}$$

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

□

Strangely, the sample complexity upper bound does not depend on $\mathbb{C}$ or D but depends logarithmically on the size of $\mathcal{H}$ [Hau88].

[Hau88] Haussler, 'Quantifying inductive bias: AI learning algorithms and Valiant's learning framework'.

4.5.2 *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 $\mathcal{X}$.

$$|\mathcal{U}_d| = 2^{|\mathcal{X}|} = 2^{(2^d)} \tag{4.36}$$

$$|\mathcal{H}| \geqslant |\mathcal{U}_d| \tag{4.37}$$

$$|\mathcal{H}| \geqslant 2^{(2^d)} \tag{4.38}$$

From Theorem 4.2 ([Hau88], finite space, consistent case: sample complexity):

$$n \in \mathcal{O}\left(\frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})\right) \tag{4.39}$$

$$n \in \mathcal{O}\left(\frac{1}{\epsilon}(2^d \ln(2) + \ln\frac{1}{\delta})\right) \therefore \tag{4.40}$$

$$n \in \mathcal{O}\left(2^d; \frac{1}{\epsilon}; \ln\frac{1}{\delta}\right) \tag{4.41}$$

Therefore, the sample complexity is not polynomial to d, and $\mathcal{U}_d$ **is not PAC Learnable**. Moreover, the "no free lunch" theorem [WM97] 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.

[WM97] Wolpert and Macready, 'No free lunch theorems for optimization'.

4.5.3 *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 due to 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.

$$\text{for } n \to \infty,$$
$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \to \mathbb{E}(\xi), \xi_i \sim \mathrm{P}. \tag{4.42}$$

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

Based on the fact that a statistic[7] of random variables can be treated itself 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:

$$\text{for } n \to \infty,$$
$$\hat{\mathrm{R}}_{\mathrm{S}}(h) = \frac{1}{n}\sum_{i=1}^{n}(\ell(x_i, y_i, h(x_i)) \to \mathbb{E}(\ell(x, y, h(x)) = \mathrm{R}(h). \tag{4.43}$$

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

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

(Chernoff-Hoeffding's inequality)

$$\Pr\left(|\hat{\mathrm{R}}_{\mathrm{S}}(h) - \mathrm{R}(h)| \geqslant \epsilon\right) \leqslant 2e^{(-2n\epsilon^2)} \tag{4.44}$$

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 of 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 a hypothesis from the space using training data, the bound still holds for any hypothesis took 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:

$$\begin{aligned}\Pr\left(|\hat{\mathrm{R}}_{\mathrm{S}}(h_1) - \mathrm{R}(h_1)| > \epsilon \vee\right.\\ |\hat{\mathrm{R}}_{\mathrm{S}}(h_2) - \mathrm{R}(h_2)| > \epsilon \vee \cdots\\ \left.\cdots \vee |\hat{\mathrm{R}}_{\mathrm{S}}(h_{|\mathcal{H}|}) - \mathrm{R}(h_{|\mathcal{H}|}) > \epsilon\right) \leqslant \sum^{|\mathcal{H}|} 2e^{(-2n\epsilon^2)}\end{aligned} \tag{4.45}$$

$$\therefore \Pr\left[\exists h \in \mathcal{H} : \left|\hat{R}_S(h) - R(h)\right| > \epsilon\right] \leqslant \sum^{|\mathcal{H}|} 2e^{(-2n\epsilon^2)} \tag{4.46}$$

$$\Pr\left[\exists h \in \mathcal{H} : \left|\hat{R}_S(h) - R(h)\right| > \epsilon\right] \leqslant 2|\mathcal{H}|e^{(-2n\epsilon^2)} \tag{4.47}$$

Theorem 4.3 ([Hau88], finite space, inconsistent case)**.** *Let $\mathcal{H}$ be a finite hypothesis class. Then, for any $0 < \delta < \frac{1}{2}$, with a probability at least $1 - \delta$, the following inequality holds [MRT12]:*

[MRT12] Mohri et al., *Foundations of Machine Learning*.

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

$$R(h) \leqslant \hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}} \tag{4.48}$$

Proof.

$$\Pr\left[\exists h \in \mathcal{H} : \left|\hat{R}_S(h) - R(h)\right| > \epsilon\right] < \delta \qquad \text{(from PAC)}$$

$$\Pr\left[\exists h \in \mathcal{H} : \left|\hat{R}_S(h) - R(h)\right| > \epsilon\right] \leqslant 2|\mathcal{H}|e^{(-2n\epsilon^2)} \qquad \text{(from (4.47))}$$

$$\therefore \delta > 2|\mathcal{H}|e^{(-2n\epsilon^2)} \tag{4.49}$$

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

$$e^{(-2n\epsilon^2)} = \frac{\delta}{2|\mathcal{H}|} \tag{4.50}$$

$$-2n\epsilon^2 = \ln\delta - ln2|\mathcal{H}| \tag{4.51}$$

$$\epsilon^2 = \frac{\ln|\mathcal{H}| + \ln 2 - \ln\delta}{2n} \tag{4.52}$$

$$\therefore \epsilon > 0 \rightarrow \epsilon = +\sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}} \tag{4.53}$$

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

$$\Pr\left[\exists h \in \mathcal{H} : \left(R(h) - \hat{R}_S(h)\right) > \epsilon\right] < \delta \tag{4.54}$$

$$\Pr\left[\forall h \in \mathcal{H} : \left(R(h) - \hat{R}_S(h)\right) \leqslant \epsilon\right] \geqslant 1 - \delta \tag{4.55}$$

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

$$\forall h \in \mathcal{H}, R(h) \leqslant \hat{R}_S(h) + \epsilon \tag{4.56}$$

$$\forall h \in \mathcal{H}, R(h) \leqslant \hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln 2/\delta}{2n}} \qquad \text{(from (4.53))}$$

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

□

We can rewrite Theorem 4.3 ([Hau88], finite space, inconsistent case) to bound the sample complexity:

Theorem 4.4 ([Hau88], finite space, inconsistent case: sample complexity)**.** *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 (4.53),

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

□

4.5.4 *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 give a general idea of this case.

One of the most striking insights of Vapnik and Chervonenkis is the idea of the *shattering coefficient* ($\mathcal{N}$). Let us take a look at the bound from Theorem 4.3 ([Hau88], finite space, inconsistent case):

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

The $\ln|\mathcal{H}|$ relates to d, the size of the *representation* of the hypothesis 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 $\mathrm{P}(h_j) \cap \mathrm{P}(h_k)$, $j \neq k$.

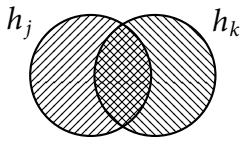


FIGURE 4.10: $\Pr(h_j) \cap \Pr(h_k)$ is summed twice in the union bound.

In reality, there are several different $h \in \mathcal{H}$ that provide the same map $x \in \mathrm{S} \to y \in \{-, +\}$. Therefore, the effective size of $\mathcal{H}$ is smaller than $|\mathcal{H}|$. Using a symmetrisation trick [VLS11, section 5.2], Vapnik and Chervonenkis showed that there are at most 2^{2n} effectively different hypotheses. In the PAC framework, $|\mathcal{Y}| = 2$, so if a pattern is a

[VLS11] Von Luxburg and Schölkopf, 'Statistical learning theory: Models, concepts, and results'.

set $\{y_1, \cdots, y_n\}$, there are $|\mathcal{H}| = 2^n$ different patterns, thus, effectively different hypotheses.This number, however, can be even smaller; for example, a certain $y_k, k < n$ can, for example, only accept a specific value, $y_k = +$.

The shattering coefficient is a growth function, *i.e.* it measures the number of effectively distinct hypotheses as the sample size n grows. It is a capacity measure of a hypothesis class. Whenever $\mathcal{N}(\mathcal{H}, n) = 2^n$, 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 4.3 ([Hau88], finite space, inconsistent case) as:

$$\forall h \in \mathcal{H},$$
$$\mathrm{R}(h) \leqslant \hat{\mathrm{R}}_S(h) + \sqrt{\frac{\ln \mathcal{N}(\mathcal{H}, n) + \ln 2/\delta}{2n}} \tag{4.57}$$

Another capacity measure is the famous VC dimension.[8]

[8] Named after Vapnik and Chervonenkis.

$$\mathrm{VC}(\mathcal{H}) = max\{n \in \mathbb{N} | \mathcal{N}(\mathcal{H}, n) = 2^n \text{ for some } \mathrm{S}_n\} \tag{4.58}$$

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

Theorem 4.5 (Vapnik, Chervonenkis, Sauer, Shelah bound)**.**

$$\textit{If } \mathrm{VC}(\mathcal{A}) = d, \forall n \geqslant 1, \qquad \mathcal{N}(\mathcal{H}, n) \leqslant \sum_{k=0}^{d} \binom{n}{k} \leqslant \left(\frac{\epsilon n}{d}\right)^d$$

4.6 MINIMUM DESCRIPTION LENGTH

Minimum Description Length (MDL) is an MLT principle proposed by Hinton and Van Camp [HVC93]. It will be presented later (Section 6.5) as it relates to Information Theory.

[HVC93] Hinton and Van Camp, 'Keeping the neural networks simple by minimizing the description length of the weights'.

4.7 PAC-BAYES

For a long time, MLT was divided between Bayesian inference and PAC learning. In 1997, Shawe-Taylor and Williamson first presented a theorem of PAC guarantees for Bayesian algorithms (algorithms that minimise the risk using a prior probability for the data and hypothesis) [STW97]. This bridge allowed tighter PAC bounds for learning algorithms that take advantage of informative priors. Here we give PAC Bayes bounds for finite hypothesis spaces (for more, see [McA99] and [McA13]).

[STW97] Shawe-Taylor and Williamson, 'A PAC analysis of a Bayesian estimator'.

[McA99] McAllester, 'Some PAC-Bayesian Theorems'.

[McA13] McAllester, 'A PAC-Bayesian Tutorial with A Dropout Bound'.

4.7.1 *PAC Bayes Guarantees for finite hypothesis spaces — consistent case*

Theorem 4.6 (Preliminary Theorem 1 [McA99]). *Let $\mathcal{H}$ be a finite hypothesis space, $\mathcal{A}$ a learning algorithm that returns a consistent hypothesis h, i.e. $\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0$, for any hypothesis h and unknown distribution $\mathrm{D} = \mathrm{P}(X, Y)$, any $|\mathrm{S}| = n : n \geqslant 1$. For any probability distribution P assigning a nonzero probability to every hypothesis in the finite hypothesis space $\mathcal{H}$, with confidence $1 - \delta$ over the selection of the sample of n instances the following holds true:*

$$Pr\left[h \in \mathrm{H} : (\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0) \wedge (\mathrm{R}_{\mathrm{D}}(h) > \epsilon)\right] \leqslant \frac{\ln \frac{1}{\mathrm{P}(h)} + \ln \frac{1}{\delta}}{n} \tag{4.59}$$

Proof. This proof is very similar to the one in Theorem 4.1 ([Hau88], finite space, consistent case). From (4.28):

$$\Pr\left[h \in \mathrm{H} : (\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0) \wedge (\mathrm{R}_{\mathrm{D}}(h) > \epsilon)\right] \leqslant (1 - \epsilon)^n, \tag{4.60}$$

But we also know that:

$$\Pr\left[h \in \mathrm{H} : (\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0) \wedge (\mathrm{R}_{\mathrm{D}}(h) > \epsilon)\right] \leqslant \mathrm{P}(h)\delta \tag{4.61}$$

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

$$e^{-\epsilon n} < \mathrm{P}(h)\delta \tag{4.62}$$

$$\epsilon < \frac{\ln \frac{1}{\mathrm{P}(h)} + \ln \frac{1}{\delta}}{n}$$

□

4.7.2 *PAC Bayes Guarantees for finite hypothesis spaces — inconsistent case*

Theorem 4.7 (Preliminary Theorem 2 [McA99]). *Let $\mathcal{H}$ be a finite hypothesis space, $\mathcal{A}$ a learning algorithm that returns a hypothesis h given a sample $|\mathrm{S}| = n : n \geqslant 1$ from the unknown distribution $\mathrm{D} = \mathrm{P}(X, Y)$. Given a probability distribution P assigning nonzero probability $\forall h \in \mathcal{H}$, with confidence $(1 - \delta)$ the following holds:*

$$\forall h \in \mathcal{H},\ \mathrm{R}(h) \leqslant \hat{\mathrm{R}}_{\mathrm{S}}(h) + \sqrt{\frac{\ln \frac{1}{\mathrm{P}(h)} + \ln \frac{2}{\delta}}{2n}} \tag{4.63}$$

Proof. As in Theorem 4.3 ([Hau88], finite space, inconsistent case), we need to apply the union bound over the Chernoff bound:

$$\Pr\left[h \in \mathrm{H} : (\hat{\mathrm{R}}_{\mathrm{S}}(h_1) = 0) \wedge (\mathrm{R}_{\mathrm{D}}(h) > \epsilon)\right] \leqslant 2e^{(-2n\epsilon^2)}, \tag{4.64}$$

But we also know that:

$$\Pr\left[h \in \mathrm{H} : (\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0) \wedge (\mathrm{R}_{\mathrm{D}}(h) > \epsilon)\right] \leqslant \mathrm{P}(h)\delta \tag{4.65}$$

$$2e^{(-2n\epsilon^2)} < \mathrm{P}(h)\delta \tag{4.66}$$

$$\epsilon < \sqrt{\frac{\ln \frac{1}{\mathrm{P}(h)} + \ln \frac{2}{\delta}}{2n}}$$

□

4.8 CRITIQUES ON MLT

This dissertation aims to present an emergent new theory for understanding Deep Learning. In this context, we should first ask ourselves: **Is anything wrong with the current MLT? Do we really need a new theory?**

Truth be told: we did not cover *current* MLT in this chapter which aimed to be an introductory overview of the subject. There are many topics in active development beyond what was presented here: Structural Risk Minimisation, Rademacher complexity, Uniform Stability, for example.

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 to the case of Deep Learning.

4.8.1 *General critiques*

No assumption on $\mathrm{D} = \mathrm{P}(X, Y)$ (see 7.2.1, assumption I): One of the assumptions of classical learning theory is that "there are no assumptions on $\mathrm{D} = \mathrm{P}(X, Y)$". Although this assumption means that MLT bounds guarantee approximation to any arbitrary distribution; distributions of practical interest are the ones found in Nature. These practical distributions have some peculiar characteristics that physicists know about [LTR17]: Low polynomial order, locality, symmetry, among others.

[LTR17] Lin et al., 'Why does deep and cheap learning work so well?'.

Absence of the notion of "time"(see 7.2.1, assumption III): One of MLT assumptions on $\mathrm{P}(X, Y)$ 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 [MP18].

[MP18] Mello and Ponti, *Machine learning: a practical approach on the statistical learning theory*.

Identically Independent sampling (see 7.2.1, assumption IV): One of the assumptions of Machine Learning is that the datasets are

sampled i.i.d. This sampling assumption is often violated in practice; for example, a machine learning medical application may use data from one hospital to train a model that will be applied worldwide.

The violations are, of course, of practical reasons. However, up to what point can we say that a particular dataset is i.i.d.? Let us think over the problem of facial recognition. Taking photos at random in a university is not i.i.d because the people that goes to the university is a limited set of the whole population. If we use random images on the Internet, we may only get the kind of picture people chose to display, a bias of intention. There is always some bias in any dataset: a selection, intention bias or technical bias (due to the image capture device).

ARBITRARY LOSS METRICS In MLT learning setting, the choice of the loss function is arbitrary, which curbs any objective, metric-independent interpretation of the results.

BLACK-BOX ANALYSIS In MLT, the model is treated as a black-box [AB16] (as cited by [ST17]), *i.e.* the analysis is based only on the input and the output of the model.

[AB16] Alain and Bengio, 'Understanding intermediate layers using linear classifier probes'.

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

4.8.2 *In specific for Deep Learning*

VACUOUS BOUNDS Machine Learning Theory cannot explain deep neural networks generalisation performance. According to MLT, the deep learning generalisation gap is in $\mathcal{O}(|\theta| \log |\theta|)$, where $|\theta|$ is the number of parameters of the network [KT08]. These bounds are vacuous by orders of magnitudes [Zho+19; Zha+16]. However, deeper and larger networks consistently show better generalisation performance than smaller ones.

[KT08] Kakade and Tewari, *VC Dimension of Multilayer Neural Networks, Range Queries*. URL: https://ttic.uchicago.edu/~tewari/lectures/lecture12.pdf

[Zho+19] Zhou et al., 'Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach'.

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization.*

"INEXPLICABLE" PHENOMENA Deep Learning (DL) has several phenomena with no definitive explanation, stemming from a single narrative. For example:

- Generalisation with the addition of layers: as we explained in this chapter, the current MLT expects models with fewer parameters to generalise better; that is not what happens in DL. Moreover, Zhang et al. showed that the hypothesis space of DNN is large enough to allow convergence to random labels [Zha+16].
- Disentanglement of semantic factors: the representation of the input in deep layers usually disentangle semantic factors, *i.e.*

different semantic factors are not strongly correlated in the representation;

- Superconvergence: Smith and Topin present that overall training time can be shortened and better accuracy achieved by cyclical learning rates [ST19]. Howard and Ruder propose a slight variation of the method, slanted triangular learning rates, and achieve even better performance [HR18]. This superconvergence phenomenon is not well studied, and there are only a few conjectures on why it does happen.

- Critical Learning Periods: Achille et al. 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" [ARS17]. This finding questions the assumption that the order in which a model experiences evidence does not affect learning.

[ST19] Smith and Topin, 'Super-convergence: Very fast training of neural networks using large learning rates'.

[HR18] Howard and Ruder, 'Universal Language Model Fine-tuning for Text Classification'.
URL: http://arxiv.org/abs/1801.06146

[ARS17] Achille et al., *Critical Learning Periods in Deep Neural Networks.*

4.9 CONCLUDING REMARKS

This chapter summarises basic concepts from Machine Learning Theory (MLT). We derived some fundamental theorems of classic MLT and PAC-Bayes (Section 4.7). We formalised the learning problem setting and made explicit its assumptions (Section 4.2.2), which we will add to our list:

4.9.1 *Assumptions*

1. A definition of intelligence (Section 2.1.1)
2. Knowledge is a set of beliefs, quantifiable by real numbers and dependent on prior evidence (Section 3.1.3, Item I);
3. Assumption of the sceptical agent language (Bayesian inference):
   a) Common sense: The plausibility of compound sentences should be related by some logical function to the plausibility of the sentences that form them (Section 3.1.3, Item II).
   b) consistency (Section 3.1.3, Item III and Section 3.1.2, Item III)
   c) minimality (Section 3.1.2, Item IV)
4. MLT specific assumptions for the learning problem:

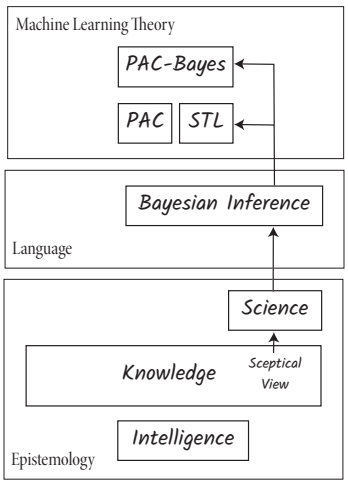


FIGURE 4.11: In this chapter we show how MLT is built from a set of specific assumptions (Item 4) using the Bayesian inference language.

a) No assumption on $\mathrm{D} = \mathrm{P}(X, Y)$;

b) $\mathrm{D} = \mathrm{P}(X, Y)$ is unknown;

c) $\mathrm{D} = \mathrm{P}(X, Y)$ is fixed: no "time" parameter.

d) Independent sampling;

e) Labels may assume non-deterministic values (h can be stochastic, but can also be deterministic);

f) Learning is an optimisation problem in the hypothesis space.

4.9.2 *Revealing the implicit assumptions*

Our derivation allowed us to expose implicit assumptions of MLT. For example, although some may argue that MLT is agnostic of a frequentist or Bayesian view, we disagree. We claim that MLT requires a Bayesian view and refer to the fact that we derived it from a Bayesian definition of Knowledge. Another point we would like to highlight is that MLT assumes that there is no importance of the order of experiences, *i.e.* it assumes **consistency** (Items 3b and 4d):

i. A belief in 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 known to be identical must be assigned the same degree of belief.

Symbolic AI guarantees that their agents follow such assumptions by construction. However, on the other hand, we know humans do not follow these assumptions, and the whole point of conceptualising rational agents was to study a simplified form of intelligence.

For humans,

i. the order in which we experience pieces of evidence matter. Humans and other animals have critical learning periods [Wie82];

ii. we forget or suppress past experiences;

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

[Wie82] Wiesel, 'Postnatal Development of the Visual Cortex and the Influence of Environment'.

What about Deep Neural Networks? There is nothing by construction that forces DNNs to be consistent. Recently, Achille et al. observed critical learning period phenomena in DNNs as well [ARS17]. Therefore, we conjecture:

[ARS17] Achille et al., *Critical Learning Periods in Deep Neural Networks.*

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

4.9.3 *On the critiques*

Most of the general critiques in Section 4.8.1 are not problems of current MLT but choices.

Specific to Deep Learning, Zhang et al. challenge current MLT concept of generalisation based on the expressivity of the model [Zha+16]. They show that the expressivity of neural networks is sufficient to fit random labels easily and even memorise an entire dataset. Randomising labels is a data transformation that does not affect the generalisation performance in current MLT generalisation bounds.

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization.*

Current MLT sample complexity and generalisation bounds, based on the size of the hypothesis space, focus research attention on models architectures. One of the strongest critiques to the theory has for a long time been the lack of non-vacuous bounds for DNNs. Recently, however, Dziugaite and Roy proved such bounds [DR17]. They did so, however, using PAC-Bayes *and* exploring the "flatness"/location of minima found by SGD, proving that at least the optimiser has a role in DL generalisation. Besides, we will show that there is an information-theoretical interpretation for the "flatness" of SGD local minima.

[DR17] Dziugaite and Roy, 'Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data'. URL: http://auai.org/uai2017/proceedings/papers/173.pdf

Nevertheless, without disregarding the immense contribution of 'Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data', the paper does not pretend to solve conceptual problems in MLT.

Understanding Deep Learning, indeed, requires rethinking generalisation. A new learning theory may make different choices and bring a new *narrative* that unifies explanations for Deep Learning phenomena. We will show that, despite its weaknesses, Information Bottleneck Theory (IBT) presents a new narrative worth exploring.

5

Information Theory

'*Only through communication can human life hold meaning.*'

—Paulo Freire

This chapter derives Shannon Information from Probability Theory, explicates some implicit assumptions in the usage of Shannon Information, and explains basic Information Theory concepts.

5.1 FROM PROBABILITY TO INFORMATION

In Section 2.3.1, we exposed that an agent updates its model of the environment from sensory data, experience. We have also shown how this update happens; a sceptical 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 more considerable change in the agent's knowledge, leading to a greater impact in her future actions. We say that those experiences are more informative.

Definition 5.1. Information is what changes belief [Sow16; Cato8].

Let us say that an agent's *prior* belief in a statement S is $\mathrm{P}(\mathrm{S})$.[1] After experiencing some evidence e, her *posterior* set of beliefs is updated to incorporate the evidence, $\mathrm{P}(\mathrm{S}|e)$.[2] The prior and the posterior are related by the product rule (Section 3.6) [Sow16]:

$$\underbrace{\mathrm{P}(\mathrm{S}|e)}_{Posterior} = \frac{\mathrm{P}(e|\mathrm{S})}{\mathrm{P}(e)} \cdot \underbrace{\mathrm{P}(\mathrm{S})}_{Prior} \tag{5.1}$$

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

[Sow16] Sowinski, 'Complexity and stability for epistemic agents: The foundations and phenomenology of configurational Entropy'.

[Cato8] Caticha, *Lectures on Probability, Entropy, and Statistical Physics*.

[1] $\mathrm{P}(\mathrm{S})$ is in fact $\mathrm{P}(\mathrm{S}|\mathrm{K})$, but we supress it to reduce the clutter.

[2] Here we are talking about *events*: $\mathrm{P}(\mathrm{S}|e)$ is a short hand for $\mathrm{P}(\mathrm{S}|\{e\} \wedge \mathrm{K})$.

This update procedure can be generalised to a set of experiences. Consider a sequence of experiences: $\mathrm{E} = \{e_t\}_0^{\mathrm{T}}$

$$p(\mathrm{S}|e_0) \to p(\mathrm{S}|e_0 \wedge e_1) \to \cdots \to p(\mathrm{S}|e_0 \wedge e_1 \wedge \cdots \wedge e_{\mathrm{T}})$$

But according to the Cox axiom Section 3.1.3 and Item III, an agent may partition her experiences in any way she chooses, and this does not affect her final belief [Sow16]. Therefore[3]:

$$\mathcal{L}(e;\mathrm{S}) = \frac{\text{Posterior}}{\text{Prior}} = \frac{\mathrm{P}(\mathrm{S}|e)}{\mathrm{P}(\mathrm{S})} = \frac{\mathrm{P}(e|\mathrm{S})}{\mathrm{P}(e)} \tag{5.2}$$

$$\mathrm{P}(\mathrm{S}|e) = \mathcal{L}(e;\mathrm{S}) \cdot \mathrm{P}(\mathrm{S}). \tag{5.3}$$

Simply by observing equation 5.2, we can conclude that if information (i) is what changes belief, information and likelihood must be related to one another:

$$i_{\mathrm{S}}(e) = f(\mathcal{L}(e;\mathrm{S})). \tag{5.4}$$

Moreover, if an experience does not change a belief ($\mathcal{L}(e;\mathrm{S}) = 1$), it contains no information: $f(1) = 0$.

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

The information gathered from independent *events* must reflect the commutativity of Cox's axiom III IIIa.

Let $\mathcal{L}_1 = \mathcal{L}(e_1;\mathrm{S})$ and $\mathcal{L}_2 = \mathcal{L}(e_2;\mathrm{S})$, information must satisfy the functional constraints [Sow16]:

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

This functional form can be solved, and its solution is [Cato8]:

$$f = \mathrm{A} \cdot \ln \mathcal{L}(e;\mathrm{S}) \therefore$$
$$i_{\mathrm{S}}(e) = \mathrm{A} \cdot \ln \mathcal{L}(e;\mathrm{S}). \tag{5.5}$$

From equations 5.5 and 5.2,

$$i_{\mathrm{S}}(e) = \mathrm{A} \cdot \ln \frac{\mathrm{P}(\mathrm{S}|e)}{p(\mathrm{S})} \tag{5.6}$$

$$i_{\mathrm{S}}(e) = \mathrm{A} \cdot \ln \mathrm{P}(\mathrm{S}|e) - \mathrm{A} \cdot \ln p(\mathrm{S}). \tag{5.7}$$

[Sow16] Sowinski, 'Complexity and stability for epistemic agents: The foundations and phenomenology of configurational Entropy'.

[3] We have already (Section 3.1.3) delved a little on the implications of the indifference to the order of evidence which is also an indifference in sequential versus simultaneous updating.

[Cato8] Caticha, *Lectures on Probability, Entropy, and Statistical Physics*.

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

$$\mathrm{A} = \frac{1}{\ln b} \rightarrow i_{\mathrm{S}}(e) = \log_b \mathrm{P}(\mathrm{S}|e) - \log_b \mathrm{P}(\mathrm{S}). \tag{5.8}$$

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

Hence, $i_{\mathrm{S}}(e) = -\Delta h(e)$, from eq. 5.7:

$$i_{\mathrm{S}}(e) = \log \mathrm{P}(\mathrm{S}|e) - \log \mathrm{P}(\mathrm{S}) \tag{5.9}$$

$$i_{\mathrm{S}}(e) = -\left[\underbrace{\left(-\log \mathrm{P}(\mathrm{S}|e)\right)}_{h(\mathrm{S}|e)} - \underbrace{\left(-\log \mathrm{P}(\mathrm{S})\right)}_{h(\mathrm{S})}\right] \tag{5.10}$$

$$i_{\mathrm{S}}(e) = -\Delta h(e). \tag{5.11}$$

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

$$h(\mathrm{S}) = -\log \mathrm{P}(\mathrm{S}) \tag{5.12}$$

is equivalent to Shannon's self information[4]:

$$\mathrm{I}[S] = -\log p(s) \tag{5.13}$$

[4] Also known as the Shannon information content of an outcome [Mac02] or Hartley's information.

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

Shannon's information can be derived from probability theory.

5.2 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*.[5] Despite his acknowledging of the influence from previous works by pioneers such Harry Nyquist and Ralph Hartley, it was Shannon's unifying vision that revolutionised communication and provided a *blueprint* for the information age [Aft+01]. His theory defines unbreachable limits, the *laws of information* [Sto15]:

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

[5] In a rare piece of collaboration, Shannon asked his lunchroom table colleagues at Bell Labs to come up with a snappier name than *binary digit*. *Bigit* was considered, but John Tukey's proposal, *bit*, was chosen [SG17].

[Aft+01] Aftab et al., *Information Theory: Information Theory and the Digital Age*.
URL: http://web.mit.edu/6.933/www/Fall2001/Shannon2.pdf

[Sto15] Stone, *Information theory: a tutorial introduction*.

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

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

The idea of transmitting information with zero error through a noisy channel is not intuitive, and its theoretical proof was an unexpected result. In the following sections, we will explain the concepts of IT that allow us to comprehend these *laws of information*.

5.2.1 *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, [Sha48]*

[Sha48] Shannon, 'A mathematical theory of communication'.

Conceptually, this setting can be explained as follows (Figure 5.1):[6]

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

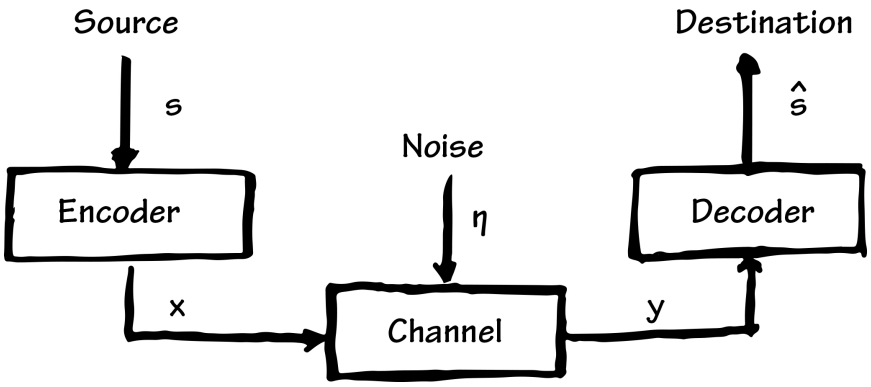


FIGURE 5.1: The communication problem setting.

The Source S:

1. selects a message s from a set of possible messages $\mathbb{A}_S$.[7]

2. The encoder x encodes the message s into a string of symbols x, the signal; and

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

3. transmits this string of inputs x through a noisy channel $p(y|x, \eta)$.

In the Destination:

1. The decoder $Y := p(y|x, \eta)$ receives a string of symbols y,
2. decodes the string y into the most probable message $\hat{s}$.

5.3 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* [Sha48, p.100]. We have already established (Section 5.1, definition 5.1) that *information is what changes belief*; thus, changes an agent's conduct. So, **communication is transmitting information**.

Noteworthy, information is independent of the *encoding* or the chosen channel. Thus, one can use any language (English, Portuguese, music, images, dance) and any transmission medium (letter, telegraphy, microwaves) 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 a phenomenon) that always sends the same message never surprises the receiver, and the message carries no information. On the contrary, a source that sends symbols at random is impossible to predict, and, therefore, every message carries maximal information.

Therefore, *information is a measure of freedom of choice in selecting the message* [SW49, p.100]. In other words, it is a measure of surprisal or uncertainty reduction.

[SW49] Shannon and Weaver, *The Mathematical Theory of Communication*.

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.

5.3.1 *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 are needed to guess the number correctly? Well, it depends on what are the allowed answers. One could ask:

- How many hundreds the number have?

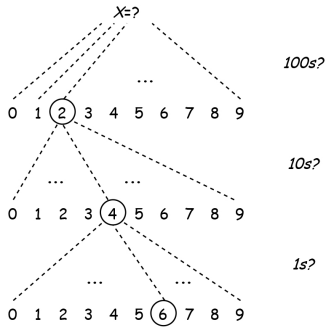


FIGURE 5.2: Branching factor of 10 to find 246.

- Then, how many tens the number have?
- Then, how many units?

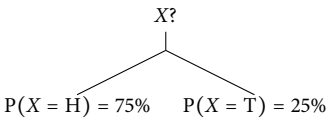
In this case, the number of questions needed is three, the height of the tree in Figure 5.2, because we allowed each answer to be a *digit*; therefore, the *branching factor b* of the decision tree was 10. It is easy to notice that the tree's height is $\log_b(1000)$.

It is now clear what Shannon meant by saying that the logarithmic function was the natural measure of information. The logarithm will give the decision tree's height (number of questions) based on the number of possible answers (the logarithm base). 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 that 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, which turns *heads* 75% of the time.



Here, we expect 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 turns *tails*, our certainty reaches 100%, growing by a factor of $\frac{1}{0.25} = 4$. So it is reasonable to think that our uncertainty of the *tails* outcome decreased 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 decreased by a factor of $\frac{1}{0.75} \approx 1.\overline{3}$. How do we transform this uncertainty reduction factor[8] to a measure in bits? In other words, how do we measure in bits the information gained by unveiling an outcome?

[8] MacKay call this factor *Occam's factor* [Mac02].

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

2^1 factor = 1 bit
2^2 factor = 2 bits
...
2^n factor = n bits
$\therefore x$ factor = $\log_2(x)$ bits

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

If the factor is a measure of the reduction in freedom of choice, the factor is the information gained by knowing the *experiment's* outcome.

Thus, this factor is known as **self-information** or information content of an outcome[9]:

Definition 5.2. The **information content, self-information, surprisal**, or **Shannon information** of a particular outcome x of an *experiment* is defined as:

$$\mathrm{I}[x] = h[x] = -\log p(x) \tag{5.14}$$

(information content of outcome)

As we already had derived in Section 5.1.

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

5.3.2 *Entropy*

In practice, however, we are not usually interested in the information of a particular outcome, but in how surprised, on average, we will expect to be with the entire set of possible outcomes.

Definition 5.3. The entropy $\mathrm{H}[X]$ of a random variable X is defined to be the average Shannon information content of its possible outcomes:

$$\mathrm{H}[X] \triangleq \mathbb{E}_p \frac{1}{\log p(x)} = -\sum_{x \in \mathbb{A}_X} p(x) \log p(x) \text{ bits/symbol.} \tag{5.15}$$

Entropy can be seen in two ways[10]:

1. as the quantity of information "produced" by the source [SW49, p.18].
2. as a measure of *uncertainty* or lack of pattern.

[10] We will constrain our explanations of Information Theory to the discrete case. It can be argued that if we are interested in models that computers will use, some quantisation will always happen.

[SW49] Shannon and Weaver, *The Mathematical Theory of Communication*.

Average information shares the same definition as Entropy; therefore, to know whether a quantity is information or Entropy depends on whether it is given or taken [Sto15]. In other words, uncertainty reduced is information gained, and vice-versa. If a random variable X is very uncertain, it has high Entropy. If we are told the outcome of the variable $X = x_j$, we have been given information equal to the uncertainty we had. Thus, receiving an amount of information is equivalent to having the same amount of Entropy taken away.

[Sto15] Stone, *Information theory: a tutorial introduction*.

5.4 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" [SW49, p.10].

A mathematical model that follows this description is known as a *stochastic process*. A stochastic process can represent any discrete source. "Conversely, any stochastic process may be considered a discrete source" [SW49].

[SW49] Shannon and Weaver, *The Mathematical Theory of Communication*.

Definition 5.4. A **stochastic (or random) process** is a set of random variables indexed by a variable $i \in \mathbb{N}$ (usually representing time):

$$S_i, i \in \mathbb{N} \tag{5.16}$$

(Stochastic Process)

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

$$R_S \triangleq \frac{H[S]}{T_S} \frac{\text{bits}}{\text{second}} \tag{5.17}$$

where T_S 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:

$$R_S = H[S] \text{ bits/symbol} \quad \text{or} \quad H[S] \text{ bits/transmission} \tag{5.18}$$

5.4.1 *Markov chains*

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

Definition 5.5. An order-k **Markov chain** is a stochastic process that satisfies the following property:

$$P(S_i|S_{i-1}, S_{i-2}, \cdots, S_{i-k}) = P(S_i|S_{i-1}, S_{i-2}, \cdots, S_1) \tag{5.19}$$

The **ergodic** property means statistical homogeneity [SW49]: its statistical properties can be deduced from a single, sufficiently long, random sample of the process.

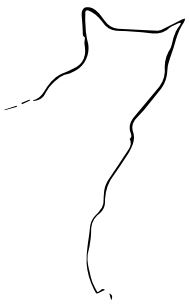
An order-k ergodic Markov chain is a process with a memory of k states. By modelling the source as an ergodic Markov chain, Shannon showed that his theory not only works for phenomena that can be modelled as i.i.d. random variables. The source can behave like a chain of random variables $\{S\}$, each 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.

5.5 DATA COMPRESSION: ENCODER/DECODER

An encoder transforms information into data. For example, the same information can be transformed into an audio file with spoken English, a piece of writing in Portuguese, or even an image. These encodings represent the information uniquely and differ in the amount of data (*bits*) they use (Figures 5.3a to 5.3c).

(a) A 360 kB PNG colored image of a cat.

(b) A 27 kB JPG grayscale image of a cat.



(c) A 4.9 kB SVG duotone image of a cat.

FIGURE 5.3: Different representations of a cat and their encoding sizes in bits.

‘What’s in a name?
That which we call a rose,
by any other word
would smell as sweet.’
— William Shakespeare,
Romeo and Juliet (act.2, sce.2)

An analogy with natural languages can better explain this idea. Languages encode ideas into words in different ways. For example, while in English *“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 [Zas+18].

[Zas+18] Zaslavsky et al., ‘Efficient compression in color naming and its evolution’.

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, characters, or size in *bits*:

$$S^n = \{S_1, \cdots, S_n\} \xrightarrow{encoding:\ \mathrm{X}(S)} \{X_1, \cdots, X_k\} = X^k. \tag{5.20}$$

$$X^k = \{X_1, \cdots, X_k\} \xrightarrow{decoding:\ \mathrm{X}^{-1}(X)} \{S_1, \cdots, S_n\} = S^n. \tag{5.21}$$

Besides, some symbols are more important in a message: “Mst nglsh spkrs wll ndrstnd ths phrs wtht vwls[11]”. Here we created *codewords* for words in English that a receiver can understand by the context (and certainly if she has a *codebook*[12]).

[11] “Most English speakers will understand this phrase without vowels”.

Shannon’s source coding theorem is about encoding messages efficiently, a form of data compression [Sto15]. Here we present some definitions that will help us understand the theorem later.

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

[Sto15] Stone, *Information theory: a tutorial introduction*.

Definition 5.6. A **(n, k) block code**, also known as a codebook, is a set of n codewords represented by a sequence of k bits:

$$\{X^k(1), X^k(2), ..., X^k(n)\},\ X^k(i) \in \mathbb{A}_X^k,\ n \in \mathbb{N}. \tag{5.22}$$

Definition 5.7. Let S^n be a block of n random variables, representing consecutive symbols $S_i \in \mathbb{A}_S$ emitted by the source. A **binary block encoder** X is a function:

$$X: \mathbb{A}_S^n \to \{0,1\}^k \tag{5.23}$$

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:

$$X(S^n) = \{x_1, \cdots, x_k\} = \boldsymbol{x} \in \{0,1\}^k \tag{5.24}$$

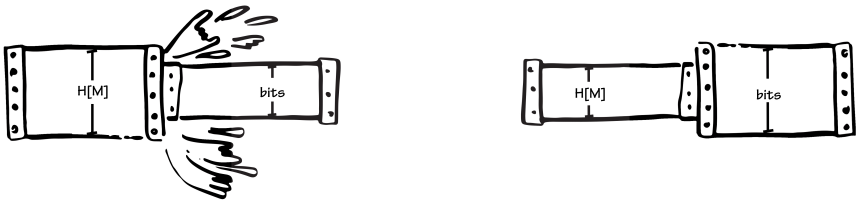
Definition 5.8. The rate $\mathbb{R}_X$ of a binary block encoder is:

$$\mathbb{R}_X = \mathbb{R}_{(n,k)} = \frac{\log|\mathbb{A}_S^n|}{k} = \frac{n}{k}\log|\mathbb{A}_S| \ \frac{\text{bits}}{\text{symbol}} \tag{5.25}$$

Shannon's source coding theorem (Section 5.5.6) is essentially about data compression. The encoding process yields inputs with a specific distribution $\mathrm{P}(X)$. The shape of this distribution[13] determines its entropy $\mathrm{H}[X]$ and, therefore, how much information each input carries [Sto15].

[13] The relationship between information (entropy) and the shape of the distribution is crucial for the IBT perspective.

[Sto15] Stone, *Information theory: a tutorial introduction*.



(a) Loss of information.

(b) Waste of resources (bits).

FIGURE 5.4: Entropy of the source vs. coding capacity.

Shannon proved a relation between the source's entropy and its optimal encoding (this relation will be shown in Section 5.5.6). The source's entropy 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 Section 5.5). 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 fewer bits than the entropy to encode it, we lose information (see Figure 5.4a). Conversely, if we use more bits than the entropy, we are wasting resources (see Figure 5.4b).

5.5.1 *An encoding example*

Let us use an example to illustrate better this crucial concept in IT[14].

[14] This example is inspired by *A Short Introduction to Entropy, Cross-Entropy and KL-Divergence* [Gé18]

Imagine building a weather station that sends the moment weather condition to a distant control room. Also, there are eight weather conditions in which we are interested. In this case, a message transmits one symbol from $\mathbb{A}_S$.

$$\mathbb{A}_S = \{w_0, w_1, w_2, w_3, w_4, w_5, w_6, w_7\} \tag{5.26}$$

How can we encode these weather conditions?

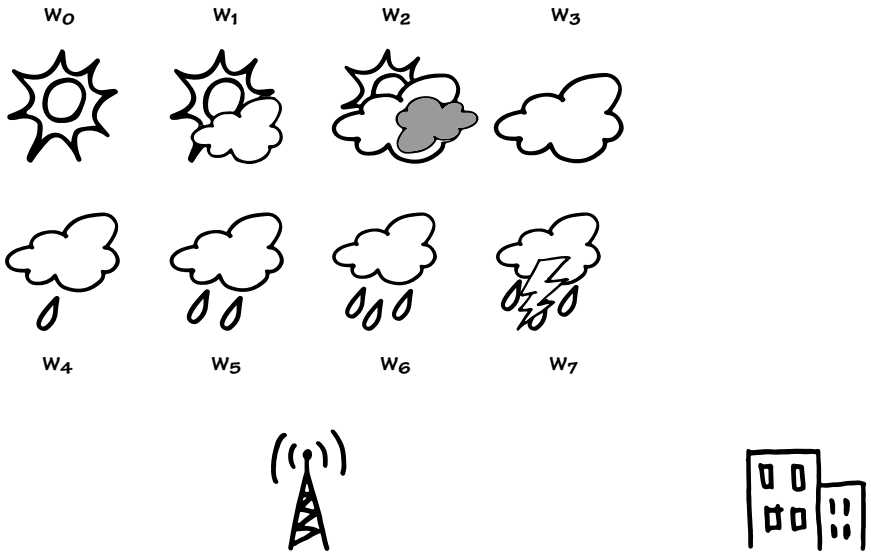


FIGURE 5.5: A weather station. Inspired by [Gé18]

5.5.2 *Raw bit content*

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

$$\begin{aligned}\mathbb{A}_X = \{&x_0 = 000, x_1 = 001, x_2 = 010, x_3 = 011,\\ &x_4 = 100, x_5 = 101, x_6 = 110, x_7 = 111\}\end{aligned} \tag{5.27}$$

This encoding provides a model of the source that has maximum entropy (all outcomes are equiprobable, thus have the same encoding size)[15]:

$$p(x_i) = \frac{1}{|\mathbb{A}_X|}, \forall i \in [0, 7] \tag{5.28}$$

$$\mathrm{H}[X] = -\sum^{|\mathbb{A}_X|} \frac{1}{|\mathbb{A}_X|} \log \frac{1}{|\mathbb{A}_X|} \tag{5.29}$$

$$= \log |\mathbb{A}_X|. \tag{5.30}$$

Is this a good encoding?

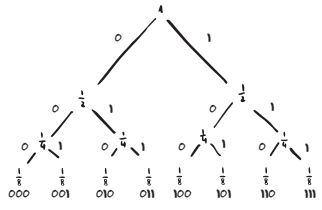


FIGURE 5.6: Largest encoding = Maximum entropy.

[15] The probability distribution that produces maximum entropy is the *uniform distribution* (Section 3.11.2)

5.5.3 *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 that conveys this information and has maximum Entropy, *i.e.* it makes no further assumptions. This maximally entropic model has the worst-case scenario for the average number of questions needed to find out which outcome is the right one:

$$\begin{aligned} \mathrm{P}(S) = \{&p_0 = \tfrac{1}{8}, p_1 = \tfrac{1}{8}, p_2 = \tfrac{1}{8}, p_3 = \tfrac{1}{8}, \\ &p_4 = \tfrac{1}{8}, p_5 = \tfrac{1}{8}, p_6 = \tfrac{1}{8}, p_7 = \tfrac{1}{8}\} \end{aligned} \tag{5.31}$$

In this case, that encoding (5.27) is indeed a good option. Notice that the encoding process yields a specific distribution $\mathrm{P}(X)$, which determines its entropy $\mathrm{H}[X]$ and, therefore, how much information per symbol it carries [Sto15]. The maximum entropy is obtained with this equiprobable distribution, the *uniform distribution* (Section 3.11.2).

[Sto15] Stone, *Information theory: a tutorial introduction*.

Let us assume now that another information about the source is given. The weather station is in the Atacama desert, and $\mathrm{P}(S') = \{p_0 = 75\%, p_1 = 10\%, p_2 = 5\%, p_3 = 1\%, p_4 = 1\%, p_5 = 1\%, p_6 = 1\%, p_7 = 1\%\}$. With this new information about the source. Can we do better? Sure.

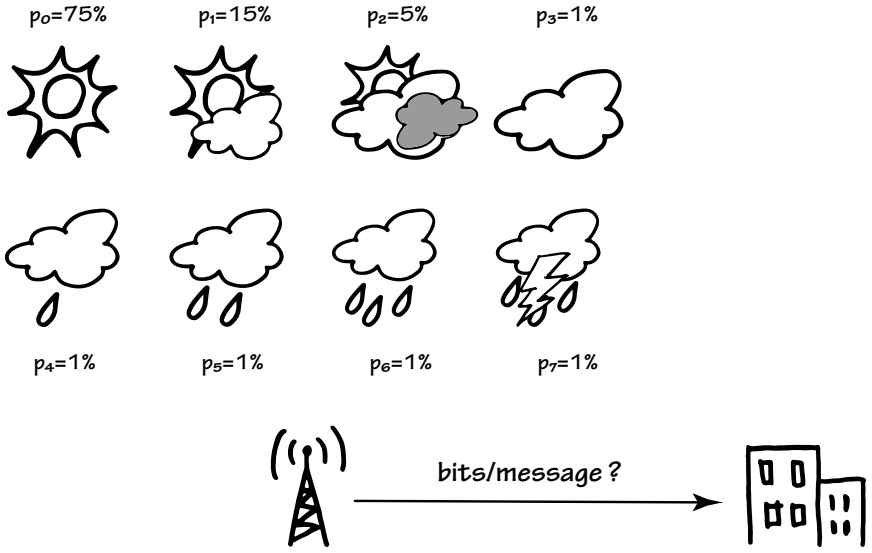


FIGURE 5.7: A weather station in the Atacama. Inspired by [Gé18]

First, let us calculate the lower bound (maximum efficiency) of the bits/symbol rate of the source encoding, $\mathbb{R}_X = \mathrm{H}[S']$:

$$\begin{aligned} \mathrm{H}[S'] &= 0.75 \log \frac{1}{0.75} + 0.15 \log \frac{1}{0.15} + 0.05 \log \frac{1}{0.05} + 5\left(0.01 \log \frac{1}{0.01}\right) \\ &\approx 1 \frac{\text{bits}}{\text{symbol}} \end{aligned} \tag{5.32}$$

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 5.6)[16]:

$$\begin{aligned}\mathbb{A}_{X'} = \{&x'_0 = 0, x'_1 = 10, x'_2 = 110, x'_3 = 11100,\\ &x'_4 = 111010, x'_5 = 111011, x'_6 = 11110, x'_7 = 11111\}\end{aligned} \tag{5.33}$$

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

$$\begin{aligned}&0.75 \cdot 1 + 0.15 \cdot 2 + 0.05 \cdot 3 + 0.03 \cdot 5 + 0.02 \cdot 6\\ &\approx 1.5 \frac{\text{bits}}{\text{symbol}}\end{aligned} \tag{5.34}$$

[16] 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.

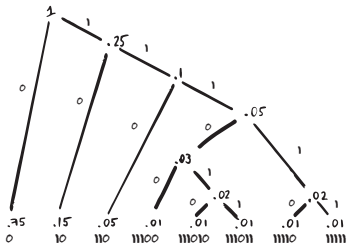


FIGURE 5.8: The probability distribution of the source determines an encoding.

5.5.4 *Cross-Entropy*

This average encoding size per message symbol 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 ($p = q$), the Cross-Entropy is equal to the Entropy $H_{p,q} = H_p$. If not (as it is in most cases), $H_{p,q} > H_p$.

In our the 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 the Atacama to London, where the probability distribution of the weather is $P(S'') = \{p_0 = 5\%, p_1 = 5\%, p_2 = 10\%, p_3 = 15\%, p_4 = 15\%, p_5 = 20\%, p_6 = 20\%, p_7 = 10\%\} \therefore H[S''] \approx 2.8$, and keep using the same encoding. The encoding will be much less efficient. The average size of a message symbol in this situation is:

$$H_{p,q}[S''],\ p = P(S''), q = P(X') \tag{5.35}$$

$$\begin{aligned}&= 0.05 \cdot 1 + 0.05 \cdot 2 + 0.1 \cdot 3 + 0.45 \cdot 5 + 0.35 \cdot 6\\ &\approx 4.8 \frac{\text{bits}}{\text{symbol}}\end{aligned} \tag{5.36}$$

The efficiency of the encoding is $2.8/4.8 = 58.33\%$.

Definition 5.9. Cross-entropy is the average number of bits needed to encode data coming from a source S with distribution $p(s)$ when using model $q(s)$.

$$H_{p,q}[S] = -\sum_{s \in \mathbb{A}_S} p(s) \log q(s) \tag{5.37}$$

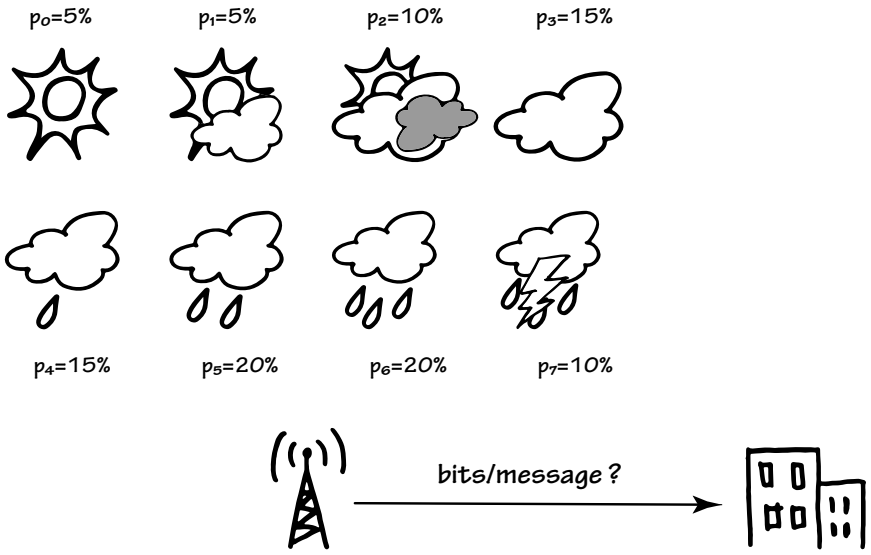


FIGURE 5.9: The the Atacama's weather station in London. Inspired by [Gé18]

5.5.5 *KL Divergence (or Relative Entropy)*

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

Definition 5.10. The **relative entropy or Kullback–Leibler divergence** between two probability distributions $p(s)$ and $q(s)$ that are defined over the same alphabet $\mathbb{A}_S$ is:

$$\mathrm{D}_{\mathrm{KL}}(p||q) = \sum_s p(s) \log \frac{p(s)}{q(s)} = \mathbb{E}_S \log \frac{p}{q} \tag{5.38}$$

$$\mathrm{D}_{\mathrm{KL}}(p||q) = \mathrm{H}_{p,q}[S] - \mathrm{H}_p[S] \tag{5.39}$$

In our example:

$$\mathrm{D}_{\mathrm{KL}}(p_{\text{the Atacama}}||q_{\text{London}}) = \underbrace{\mathrm{H}_{p,q}[S'']}_{\approx 4.8} - \underbrace{\mathrm{H}_p[S'']}_{\approx 2.8} \approx 2 \frac{\text{bits}}{\text{symbol}} \tag{5.40}$$

5.5.6 *Shannon's source encoding theorem*

Now that we understand how the source encoding works, let us take a moment to appreciate the geniality 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 5.1 (Shannon's 1st Law)**.** *The optimal binary encoding* $X^k = (\mathrm{X}_1, \cdots, \mathrm{X}_k)$, $X_i \in \{0, 1\}$, *of a n-symbols message* $S^n = (S_1, \cdots, S_n)$, *where* $S_i \in \mathbb{A}_S$ *are i.i.d.* $\sim p(s)$ *has an expected size* $k \approx n\mathrm{H}[S]$ *for sufficiently large* n.

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:

$$k \leqslant \log\lceil |S^n| \rceil = \log\lceil 2^{n \log |\mathbb{A}_S|} \rceil = n \log |\mathbb{A}_S| + 1 \text{ bits} \tag{5.41}$$

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 are most probable. So let us use this property of Nature.

An ~~obsessive~~ observant reader may have noticed that we are here considering the source as an i.i.d. stochastic process, instead of a stationary ergodic process. This is the same proof stated by Shannon [Sha48] and others [CT06; Mac02]. A proof for ergodic finite alphabet sources can be found in 'The Basic Theorems of Information Theory' [McM53].

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 5.10.

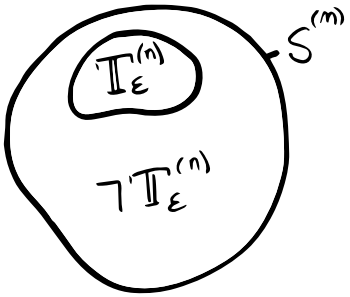


FIGURE 5.10: The typical set of sequences S^n.

Definition 5.11. 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:

$$\begin{cases} \mathrm{P}(\mathbb{T}_\epsilon^{(n)}) = \sum_{S^n \in \mathbb{T}_\epsilon^{(n)}} \mathrm{P}(S^n) > 1 - \epsilon, \text{ for sufficiently large } n \\ \mathrm{P}(S^{(n)} \in \mathbb{T}_\epsilon^{(n)}) \approx p(s_i), \forall i. \end{cases} \tag{5.42}$$

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 sure to belong to the typical set $\mathbb{T}_\epsilon^{(n)}$, if n is large, and the probability of any outcome is almost the same.

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

We know that:

$$|\mathbb{T}_\epsilon^{(n)}| \ll |\neg\mathbb{T}_\epsilon^{(n)}| < |S^n|, \tag{5.43}$$

$$\mathrm{P}(\mathbb{T}_\epsilon^{(n)}) \gg \mathrm{P}(\neg\mathbb{T}_\epsilon^{(n)}), \tag{5.44}$$

$$\mathbb{E}(k) = \lceil \mathrm{P}(\mathbb{T}_\epsilon^{(n)}) \log |\mathbb{T}_\epsilon^{(n)}| + \mathrm{P}(\neg\mathbb{T}_\epsilon^{(n)}) \log |\neg\mathbb{T}_\epsilon^{(n)}| \rceil. \tag{5.45}$$

Therefore, from (5.41) we can predict that:

$$\mathbb{E}(k) \ll n \log |\mathbb{A}_S| + 1 \text{ bits} \tag{5.46}$$

Now, we need to find $|\mathbb{T}_\epsilon^{(n)}|$. For this, we will use the Asymptotic Equipartition Property (AEP), formalised bellow [CT06]:

[CT06] Cover and Thomas, *Elements of Information Theory*.

Theorem 5.2 (AEP)**.** *If $S_1, \cdots, S_n$ are i.i.d. sampled from the same distribution $p(s)$, then:*

$$-\frac{1}{n} \log \mathrm{P}(S_1, \cdots, S_n) \to \mathrm{H}[S] \textit{ in probability.} \tag{5.47}$$

Proof. From the theorem definition, S_i are independent. Then from the Product Rule (eq. 3.11):

$$-\frac{1}{n}\sum_{i=1}^{n}\log \mathrm{P}(\underbrace{S_1,\cdots,S_n}_{S^n}) \overset{\text{eq. 3.11}}{=} -\frac{1}{n}\log(\prod_{i=1}^{n}\mathrm{P}(S_i))^{p(s)} \tag{5.48}$$

$$= \frac{1}{n}\sum_{i=1}^{n} -\log p(s) \tag{5.49}$$

From the weak law of large numbers:

$$n \to \infty,\ \frac{1}{n}\sum_{i=1}^{n}\xi_i \to \mathbb{E}(\xi) \tag{5.50}$$

Therefore, using the fact that a statistic of a random variable is a random variable, let $\xi = -\log \mathrm{P}(S_i)$ [CT06] and using (5.15) and (5.50):

[CT06] Cover and Thomas, *Elements of Information Theory*.

$$n \to \infty,$$

$$\frac{1}{n}\sum_{i=1}^{n}(-\log \mathrm{P}(S_i)) \to \underbrace{\mathbb{E}_p(-\log p(s))}_{\mathrm{H}[S]} \tag{5.51}$$

$$\therefore -\frac{1}{n}\log \mathrm{P}(S^n) \to \mathrm{H}[S] \quad \square \tag{5.52}$$

$\square$

Now that we proved Theorem 5.2 (AEP), let us use it to define $|\mathbb{T}_\epsilon^{(n)}|$:

$$-\frac{1}{n}\log \mathrm{P}(S^n) \to \mathrm{H}[S] \text{ in probability} \tag{5.53}$$

$$\mathrm{P}(S^n) \to 2^{-n(\mathrm{H}[S])} \therefore \tag{5.54}$$

$$2^{-n(\mathrm{H}[S]+\epsilon)} \leqslant \mathrm{P}(S^n) \leqslant 2^{-n(\mathrm{H}[S]-\epsilon)} \text{ in probability} \tag{5.55}$$

We also know that:

$$1 = \sum_{S^n}\mathrm{P}(S^n) \tag{5.56}$$

$$1 \geqslant \sum_{S^n \epsilon \mathbb{T}_\epsilon^{(n)}}\mathrm{P}(S^n) \tag{5.57}$$

$$1 \geqslant |\mathbb{T}_\epsilon^{(n)}|\,\mathrm{P}(S^n) \tag{5.58}$$

From (5.55):

$$1 \geqslant |\mathbb{T}_\epsilon^{(n)}|\,2^{-n(\mathrm{H}[S]+\epsilon)} \tag{5.59}$$

$$\therefore |\mathbb{T}_\epsilon^{(n)}| \leqslant 2^{n(\mathrm{H}[S]+\epsilon)} \tag{5.60}$$

This upper bound to $|\mathbb{T}_\epsilon^{(n)}|$ is all we need to prove Theorem 5.1 (Shannon's 1st Law).

$$\begin{aligned}
\mathbb{E}(k) &= \lceil P(\mathbb{T}_\epsilon^{(n)}) \log |\mathbb{T}_\epsilon^{(n)}| \\
&\quad + \cancelto{\epsilon}{P(\neg \mathbb{T}_\epsilon^{(n)})} \cancelto{|S^n| = n \log |\mathbb{A}_S|}{\log |\neg \mathbb{T}_\epsilon^{(n)}|} \rceil && (5.61)\\
&\simeq \lceil (1-\epsilon) \log 2^{n(H[S]+\epsilon)} + \cancelto{\epsilon' n}{\epsilon n \log |\mathbb{A}_S|} \rceil && (5.62)\\
&\simeq \lceil (1-\epsilon)[n(H[S]+\epsilon)] + \epsilon' n \rceil && (5.63)\\
&\simeq \lceil n(H[S] + \epsilon - \epsilon n H[S] - \epsilon^2) + n(\epsilon') \rceil && (5.64)\\
&\simeq \lceil n(H[S] + \epsilon - \cancelto{\epsilon''}{\epsilon H[S] - \epsilon^2}) + n(\epsilon') \rceil && (5.65)\\
&\simeq \lceil n(H[S] + \epsilon'' + \epsilon') \rceil = \lceil n(H[S] + \varepsilon) \rceil && (5.66)\\
&\therefore \\
\mathbb{E}(k) &\simeq n H[S] && \square
\end{aligned}$$

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.

5.5.7 *Typical Set*

We defined the typical set and discovered some of its properties in the proof of the source coding theorem, but we left one behind. We only needed the upper bound for $|\mathbb{T}_\epsilon^{(n)}|$, let us now derive its lower bound. From (5.55) and the typical set definition (5.42):

$$\sum_{S^n \epsilon \mathbb{T}_\epsilon^{(n)}} 2^{-n(H[S]-\epsilon)} \geqslant 1-\epsilon \tag{5.67}$$

$$|\mathbb{T}_\epsilon^{(n)}| 2^{-n(H[S]-\epsilon)} \geqslant 1-\epsilon \tag{5.68}$$

$$|\mathbb{T}_\epsilon^{(n)}| \geqslant (1-\epsilon) 2^{n(H[S]-\epsilon)} \tag{5.69}$$

Therefore, from (5.69) and (5.60) we can derive:

$$(1-\epsilon) 2^{n(H[S]-\epsilon)} \leqslant |\mathbb{T}_\epsilon^{(n)}| \leqslant 2^{n(H[S]+\epsilon)} \tag{5.70}$$

$$|\mathbb{T}_\epsilon^{(n)}| \to 2^{nH[S]} \tag{5.71}$$

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 (5.42) ;

2. elements in the typical set are nearly equiprobable (5.55);
3. the number of elements in the typical set is nearly $2^{H[S]}$ (5.71).

Going back to Theorem 5.2 (AEP):

$$\frac{1}{n}\log\left(\frac{1}{P(S^n)}\right) \to H[S]$$

$$H[S] - \epsilon \leqslant \frac{1}{n}\log\left(\frac{1}{P(S^n)}\right) \leqslant H[S] + \epsilon \tag{5.72}$$

We can think of the middle term as the Entropy of a sample size n. Thus a typical sample gives us an amount of information close to the average information from the source, $H[S]$[17].

[17] This insight reminds us of the sample complexity, discussed in Chapter 4

5.6 THE CHANNEL: DATA TRANSMISSION

The channel is simply the medium used to transmit the signal $\boldsymbol{x}$ from the encoder to the decoder[18]. 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 Section 5.5).[19]

[18] $x \mapsto y$

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

Definition 5.12. Mathematically, a **discrete channel** is the conditional probability

$$p(y|x), y \in \mathbb{A}_Y, x \in \mathbb{A}_X. \tag{5.73}$$

5.6.1 *Noiseless Channel Capacity*

Definition 5.13. The **operational capacity** of a channel is the maximum rate of bits per transmission that the medium is physically capable of transmitting. It is, in fact, just a number of bits per transmission. We can think of it as the maximum entropy it is capable of transmitting in the absence of noise:

$$\mathbb{C}_{\text{operational}} = R = \max_{p(x)} \log |\mathbb{A}_X| \text{ bits/usage.} \tag{5.74}$$

5.6.2 *The noisy channel*

[Sto15] Stone, *Information theory: a tutorial introduction*.

All practical communications, however, are noisy [Sto15]. Noise reduces the rate at which information can be communicated reliably. Shannon proved that information could be communicated, with arbitrarily small error, at a rate 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**.

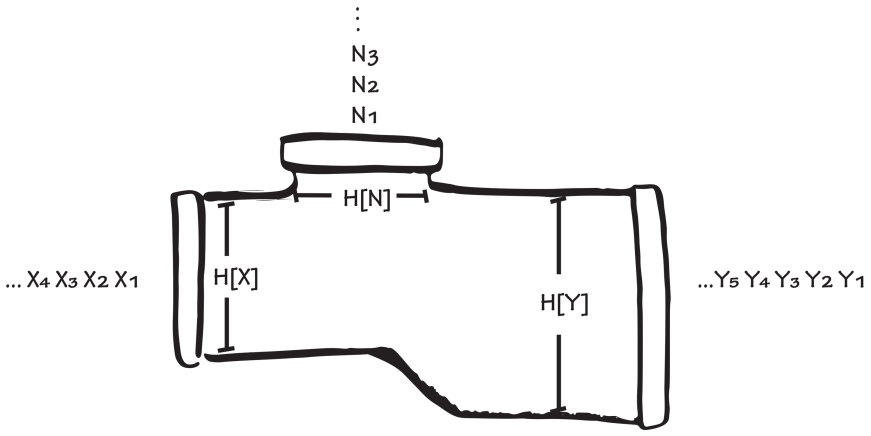


FIGURE 5.11: A noisy channel.

5.6.3 *Conditional Entropy*

The residual uncertainty we have about a random variable given that we already know the outcome of another random variable is the **conditional entropy**[20]:

[20] In the communication setting, we usually want to know the residual information in Y that is not from X, $\mathrm{H}[Y|X]$, which we call noise.

Definition 5.14. The **conditional entropy** or **equivocation** $\mathrm{H}[X|Y]$ of X given Y is:

$$\mathrm{H}[X|Y] \triangleq \sum_{y \in \mathbb{A}_Y} p(y) \left[\sum_{x \in \mathbb{A}_X} p(x|y) \log \frac{1}{p(x|y)} \right] \tag{5.75}$$

$$= - \sum_{xy \in \mathbb{A}_X \mathbb{A}_Y} p(x,y) \log p(x|y) \tag{5.76}$$

5.6.4 *Joint Entropy*

We have defined the entropy of a single random variable in (5.15). 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 [CT06, p.15].

[CT06] Cover and Thomas, *Elements of Information Theory*.

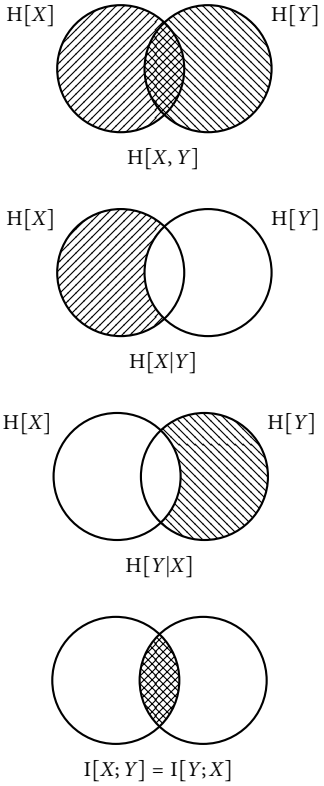
Definition 5.15. The **joint entropy** $\mathrm{H}[X, Y]$ of a pair of discrete random variables (X, Y) with joint distribution $p(x, y)$ is defined as:

$$\mathrm{H}[\mathrm{X}, \mathrm{Y}] \triangleq -\mathbb{E} \log \mathrm{P}(X, Y) \tag{5.77}$$

$$= - \sum_{x \in \mathbb{A}_x} \sum_{y \in \mathbb{A}_y} p(x,y) \log p(x,y). \tag{5.78}$$

5.6.5 *Mutual Information*

Definition 5.16. The **mutual information** $\mathrm{I}[X; Y]$ between two variables, such as a channel input X and output Y, is the amount of information obtained about one random variable through observing



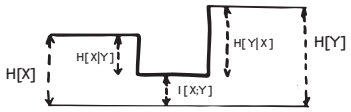


FIGURE 5.12: Relationship between information measures in a channel.

the other random variable.

$$\mathrm{I}[X;Y] = \sum_i \sum_j p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \text{ bits} \tag{5.79}$$
$$= \mathrm{H}[X] - \mathrm{H}[X|Y] \tag{5.80}$$
$$= \mathrm{H}[Y] - \mathrm{H}[Y|X] \tag{5.81}$$
$$= \mathrm{H}[X] + \mathrm{H}[Y] - \mathrm{H}[X, Y] \tag{5.82}$$
$$= \mathrm{H}[X, Y] - [\mathrm{H}[X|Y] + \mathrm{H}[Y|X]] \text{ bits} \tag{5.83}$$

For a visual understanding of these measures, see Figure 5.12. 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:

$$\mathrm{I}[X;Y] = \mathrm{D}_{\mathrm{KL}}(p(x,y) \| p(x)p(y)). \tag{5.84}$$

5.6.6 *Data Processing Inequality*

We cannot increase information by applying a deterministic function to the data, nor decrease information if the deterministic function is invertible.

Theorem 5.3 (DPI). *Let three random variables form the Markov chain $X \to Y \to Z$, implying:*

$$p(x, y, z) = p(z|y)p(y|x)p(x). \tag{5.85}$$

No processing of Y, deterministic or random, can increase the information that Y contains about X:

$$\mathrm{I}[X;Y] \geqslant \mathrm{I}[X;Z] \tag{5.86}$$

Proof. We refer to [CT06, th.2.8.1] for proof. □

[CT06] Cover and Thomas, *Elements of Information Theory*.

Theorem 5.4 (reparametrisation invariance (RI)). *Let $X \to Y \to Z$ form a Markov Chain, then functions of the data Y cannot increase the information about X,* i.e. $\mathrm{I}[X;Y] \geqslant \mathrm{I}[X;g(Y)]$.

Proof. $Z = g(Y) \therefore \mathrm{I}[X;g(Y)] = \mathrm{I}[X;Z]$. By the Data Processing Inequality (DPI) property:

$$\mathrm{I}[X;Y] \geqslant \mathrm{I}[X;Z] \tag{5.87}$$
$$\mathrm{I}[X;Y] \geqslant \mathrm{I}[X;g(Y)] \tag{5.88}$$

□

5.6.7 *Noisy channel capacity*

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

$$\begin{aligned} \mathrm{I}[X;Y] &= \mathrm{H}[Y] - \mathrm{H}[Y|X] && (5.89)\\ &= \mathrm{H}[Y] - \mathrm{H}[(X+\eta)|X]. && (5.90)\end{aligned}$$

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

$$\mathrm{I}[X;Y] = \mathrm{H}[Y] - \mathrm{H}[\eta|X] \qquad (5.91)$$

By definition, η and X are independent, therefore:

$$\begin{aligned} \mathrm{I}[X;Y] &= \mathrm{H}[Y] - \mathrm{H}[\eta] && (\text{from } (5.89))\\ \therefore \mathrm{H}[Y|X] &= \mathrm{H}[\eta] && (5.92)\end{aligned}$$

Definition 5.17. The **information capacity** or *effective capacity* of a noisy channel is defined as:

$$\begin{aligned} \mathrm{C} &= \max_{p(x)} \mathrm{I}[X,Y] && (5.93)\\ &= \max_{p(x)} (\mathrm{H}[Y] - \mathrm{H}[Y|X]) \text{ bits/transmission.} && (5.94)\\ &= \max_{p(x)} (\mathrm{H}[X] - \mathrm{H}[X|Y]) \text{ bits/transmission.} && (5.95)\end{aligned}$$

The information capacity can be derived theorem from Shannon's noisy channel theorem (5.7).

5.7 SHANNON'S NOISY CHANNEL THEOREM

In his second and, perhaps, most crucial theorem, Shannon proved that provided $\mathrm{H}[X] \leqslant \mathrm{C}$, the average error ($\epsilon$), 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 [CT06, p. 198].

Theorem 5.5 (Shannon's 2nd Law)**.** *All rates below capacity C are achievable. Specifically, for every rate* $\mathrm{R} < \mathrm{C}$*, there exists a sequence of* $(2^{n\mathrm{R}}, n)$ *codes with maximum probability of error* $\lambda^{(n)} \to 0$*. Conversely, any sequence of* $(2^{n\mathrm{R}}, n)$ *codes with* $\lambda^{(n)} \to 0$ *must have* $\mathrm{R} \leqslant \mathrm{C}$*.*

Once again, Shannon proved with a counterintuitive argument. He demonstrates 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 [Mac02]

[Mac02] MacKay, *Information Theory, Inference, and Learning Algorithms.*

and [CT06]), let us give an intuitive explanation 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 produce disjoint output sequences.

Take $\boldsymbol{x} \sim p(X^n)$. Recall Theorem 5.1 (Shannon's 1st Law), the total number of typical output sequences $\boldsymbol{y}$ is $2^{n\mathrm{H}[Y]}$ (see Figure 5.13 (B)), all sequences being almost equiprobable. For any sequence $\boldsymbol{x}$, there are about $2^{n\mathrm{H}[Y|X]}$ probable sequences (see Figure 5.13 (A)).

Now we restrict ourselves to the subset of the typical inputs, such that the corresponding typical output sets are disjoint. We can expect the *number of non-confusable inputs* to be:

$$|\mathbb{A}^{\epsilon}_{X\to Y}| \leqslant \frac{2^{n\mathrm{H}[Y]}}{2^{n\mathrm{H}[Y|X]}} = 2^{n(\mathrm{H}[Y]-\mathrm{H}[Y|X])} = 2^{n\mathrm{I}[X;Y]} \tag{5.96}$$

The maximum value of this bound is achieved by the process X that maximises $\mathrm{I}[\mathrm{X};\mathrm{Y}]$. Therefore, $n \max_{p(x)} \mathrm{I}[\mathrm{X};\mathrm{Y}]$ is the maximum amount of bits that can be transmitted in n usages of the channel, which proves the first law of information (see Section 5.2):

$$\mathrm{C}_{\text{noisy channel}} = \max_{p(x)} \mathrm{I}[X, Y]. \tag{5.97}$$

We can rewrite (5.97) as:

$$\mathrm{C}_{\text{noisy channel}} = \max_{p(x)} (\mathrm{H}[X] - \mathrm{H}[\eta]), \tag{5.98}$$

which states that noise reduces channel capacity. So, this is also a proof for the second law of information (Section 5.2).

[CT06] Cover and Thomas, *Elements of Information Theory*.

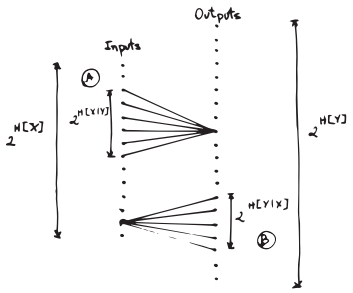


FIGURE 5.13: The need to restrict to the subset of typical inputs.

5.8 BEYOND SHANNON'S INFORMATION

Even before Shannon's 'A mathematical theory of communication', other information measures have been defined and studied. In this section we will expose two other notions of information that we will use further in the dissertation: Algorithmic information and Fisher information.

5.8.1 *Algorithmic information (Kolmogorov-Chaitin complexity)*

Developed independently by Chaitin, Solomonoff, and Kolmogorov in the 1960s, *algorithmic information* (most commonly known as

Kolmogorov complexity) of an object (*e.g.* a message) is the length of the shortest program capable of producing the object as an output [Sto15].

[Sto15] Stone, *Information theory: a tutorial introduction.*

For example, in this definition the string:
'T6ucFndKEjTyqIGYuXUKqI6fJ6HBRL'
is more complex than
'abcabcabcabcabcabcabcabcabcabc'. We can express both in the Python programming language as an example:

```
''T6ucFndKEjTyqIGYuXUKqI6fJ6HBRL''
```

versus

```
''abc''*10
```

If the object is compressable (shorter program), it has more regularity. Thus, there is a relation between complexity and compressibility.

5.8.2 *Fisher Information*

Let P_θ denote a family of parametric distributions on a space $\mathcal{X}$ with probability mass or density function given by p_θ.

Definition 5.18 (Fisher information)**.** The Fisher information $\mathrm{I}_X(\theta)$ of a random variable X w.r.t. the parameter θ is the matrix:

$$[\mathrm{I}_X(\theta)]_{ij} := \mathbb{E}_\theta\left[\nabla_{\theta_i} \log p_\theta(X) \cdot \nabla_{\theta_j} \log p_\theta(X)^\top\right] \tag{5.99}$$

$$= \mathbb{E}_\theta\left[\frac{\partial \ell}{\partial \theta_i} \cdot \frac{\partial \ell}{\partial \theta_j}^\top\right], \tag{5.100}$$

where $\ell(x|\theta) = \log p(x|\theta)$ is often called the score function.

The Fisher information measures the overall sensitivity of the functional relationship p to changes of θ by weighting the sensitivity at each potential outcome x w.r.t $p_\theta(x)$ [Ly+17]

[Ly+17] Ly et al., *A Tutorial on Fisher Information.*

A common simplification of the Fisher Information Matrix (FIM) is to reduce it to the diagonal:

$$[\mathrm{I}_X(\theta)]_i := \mathbb{E}_\theta\left[\nabla_{\theta_i} \log p_\theta(X)^2\right] \tag{5.101}$$

5.8.3 *Occam factor*

There are countless problems in science that require that given a limited dataset, preferences be assigned to alternative hypotheses of different complexities. The **Occam's razor** is the principle that states a

preference for simple theories. Although it is often advocated for aesthetic reasons, MacKay gave a Bayesian explanation for its empirical success that does not depend on any bias towards beauty [Mac02].

[Mac02] MacKay, *Information Theory, Inference, and Learning Algorithms.*

Consider evaluating the plausibility of two alternative theories $\mathcal{H}_1$ and $\mathcal{H}_2$, in the light of given evidence C (Figure 5.14). Simple models make precise predictions, while complex models are capable of making a greater variety of predictions. Hence, if $\mathcal{H}_2$ is more complex, it must spread its predictive capability more thinly over the data space D than $\mathcal{H}_1$. Thus, where the gathered data C is compatible with both theories, the simpler $\mathcal{H}_1$ will be more probable than $\mathcal{H}_2$.

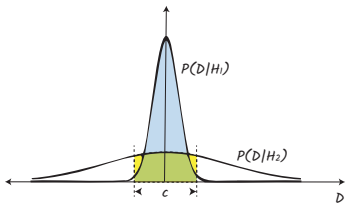


FIGURE 5.14: Comparing models $\mathcal{H}_1$ and $\mathcal{H}_2$.

$$\frac{\mathrm{P}(\mathcal{H}_1|\mathrm{D})}{\mathrm{P}(\mathcal{H}_2|\mathrm{D})} = \frac{\mathrm{P}(\mathcal{H}_1)}{\mathrm{P}(\mathcal{H}_2)} \frac{\mathrm{P}(\mathrm{D}|\mathcal{H}_1)}{\mathrm{P}(\mathrm{D}|\mathcal{H}_2)} \tag{5.102}$$

$$\therefore \mathrm{P}(\mathcal{H}_1) = \mathrm{P}(\mathcal{H}_1), \tag{5.103}$$

$$\frac{\mathrm{P}(\mathcal{H}_1|\mathrm{D})}{\mathrm{P}(\mathcal{H}_2|\mathrm{D})} = \frac{\mathrm{P}(\mathrm{D}|\mathcal{H}_1)}{\mathrm{P}(\mathrm{D}|\mathcal{H}_2)} \tag{5.104}$$

QUANTIFYING OCCAM'S RAZOR We already established that we can rank models based by evaluating the evidence $\mathrm{P}(\mathrm{D}|\mathcal{H}_i)$ (5.104):

$$\mathrm{P}(\mathrm{D}|\mathcal{H}_i) = \int \mathrm{P}(\mathrm{D}|\boldsymbol{w}, \mathcal{H}_i)\mathrm{P}(\boldsymbol{w}|\mathcal{H}_i)d\boldsymbol{w} \tag{5.105}$$

Taking for simplicity the one-dimensional case and applying Laplace's method, we can approximate the evidence by multiplying the peak of $\mathrm{P}(\mathrm{D}|\mathcal{H}_i)$ by $\sigma_{\boldsymbol{w}|\mathrm{D}}$ (approximating the shaded areas in Figure 5.14) [Mac02]:

$$\underbrace{\mathrm{P}(\mathrm{D}|\mathcal{H}_i)}_{\text{Evidence}} \simeq \underbrace{\mathrm{P}(\mathrm{D}|\boldsymbol{w}_{\mathrm{MP}}, \mathcal{H}_i)}_{\text{Best-fit likelihood}} \times \underbrace{\mathrm{P}(\boldsymbol{w}_{\mathrm{MP}}|\mathcal{H}_i)\sigma_{\boldsymbol{w}|\mathrm{D}}}_{\text{Occam's factor}} \tag{5.106}$$

The Occam's factor is the amount by which the accessible volume of $\mathcal{H}_i$'s hypothesis space collapses when data arrive. This relates to how we measure information (Section 5.3.1). **The Occam's factor log is a measure of the amount of information we gain about the model's parameters when data arrive.**

The Occam's factor is the basis of MacKay's Evidence Framework. The connection was no surprise given that we derived Information from the Bayesian interpretation of Probability (Section 5.1).

5.9 CONCLUDING REMARKS

This chapter derived the *information* measure from its definition and then summarised information-theoretical concepts.

5.9.1 *Assumptions*

1. A definition of intelligence (Section 2.1.1)
2. Knowledge is a set of beliefs, quantifiable by real numbers and dependent on prior evidence (Section 3.1.3, Item I);
3. Bayesian inference assumptions:
   a) Common sense (Section 3.1.3, Item II);
   b) Consistency (Section 3.1.3, Item III);
   c) Minimality (Section 3.1.2, Item IV).
4. MLT specific assumptions for the learning problem:
   a) No assumption on D = P(X, Y);
   b) D = P(X, Y) is unknown;
   c) D = P(X, Y) is fixed: no "time" parameter.
   d) Independent sampling;
   e) Labels may assume non-deterministic values (h can be stochastic, but can also be deterministic);
   f) Learning is an optimisation problem in the hypothesis space.
5. IT-specific assumptions:
   a) Information is what changes belief;
   b) $\mathbb{A}_S$ and $\mathbb{A}_X$ are finite sets;
   c) Sampling from an ergodic stochastic process and sampled data is typical;
   d) Labels may assume non-deterministic values (an encoder-/decoder can be stochastic or deterministic).

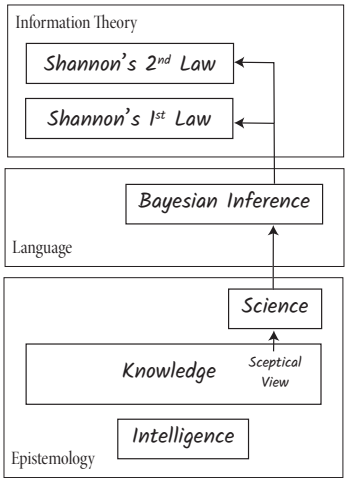


FIGURE 5.15: In this chapter we show how IT is built from a set of specific assumptions (Item 5) using the Bayesian inference language. Similarly to what was done with MLT in the last chapter.

5.9.2 *The first comparison between MLT and IT*

At this point, we have not yet expressed the Machine Learning Problem as an Information Theory problem. Still, as MLT and IT both share *Bayesian inference* as the basis they do not invalidate each other. Both may have found the same truths by different paths.

The main differences in IT from MLT assumptions are Items 5b and 5c. In Chapter 6, we will see that the first is not a problem at all. The ergodic process sampling, in its turn, is a less constrained

assumption than the i.i.d. sampling in MLT. For simplification sake, we may assume that both sample i.i.d.

Part II

INTERMEZZO


SEM AÇUC
Coca

6

Information-Theoretical Machine Learning: An Epistemology

'*Understanding is Compression*'

—Gregory Chaitin,

Meta Math! The Quest for Omega, p.65

This chapter discusses an Information-Theoretical Machine Learning (ITML)[1] perspective not specific to the Information Bottleneck (IB) Principle.

[1] We call ITML to differentiate from IBT and ITL.

6.1 LEARNING AS A CONVERSATION WITH NATURE

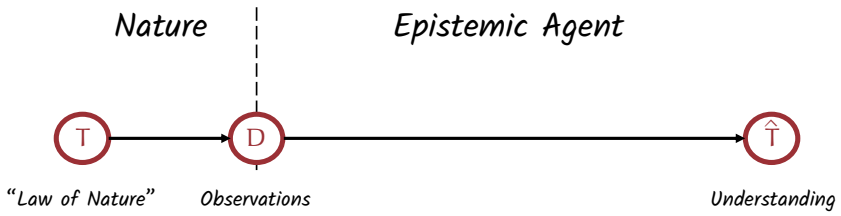


FIGURE 6.1: An understanding for a *law of nature*.

Imagine some "Law of Nature" (T)[2,3] an epistemic agent can comprehend.[4] T explains the relationship among observations in D. We can think of learning as communication between Nature and the epistemic agent.

We assume learning is possible, *i.e.* T is encoded in the observed data D. $\hat{T}$ is what the epistemic agent understand about T through D, *i.e.* a representation of T in the agent's "mind". $\hat{T}$ is the agent's

[2] T for Truth or Theorem.

[3] This chapter expands the idea of science as a conversation with Nature from [GS18]

[4] If the epistemic agent can comprehend T, $\mathrm{H}[T]$ can fit in the finite epistemic agent "mind".

understanding.

$$\hat{T} := \mathrm{U}(D) \qquad (\hat{T} \text{ is an understanding of } T \text{ through } D)$$

In this scenario, D is an **expression** of T.

$$\mathrm{E}(T) =: D \qquad (D \text{ is an expression of } T)$$

As we do not know the smallest representation size of T, $\mathrm{H}[T]$, we do not know if the $T \to D$ channel capacity ($\mathrm{C}_D = \mathrm{I}[D;T]$), is enough to noiseless transmit T through D. Therefore, we have to admit that the encoding of T into D is lossy.[5] Thus, E is stochastic, and the understanding of the agent shall stochastic as well:

$$\mathrm{E}(T) = \mathrm{P}(D|T), \tag{6.1}$$
$$\mathrm{U}(D) = \mathrm{Q}(\hat{T}|D). \tag{6.2}$$

While $\hat{T}$ is only in the epistemic agent mind, it has no practical importance for other agents. The agent will need to encode $\hat{T}$ into an agreed (n,k) language/code to communicate with other agents.

A **hypothesis** h is the epistemic agent's attempt to represent the compressed description of the observation in her mind into the agreed language X, $h := X(\hat{T})$. Without loss of generality, we can assume that any agent mind has the same size in bits and, as a consequence, $X(\hat{T})$ is a lossless encoding. Therefore,[6] $h := \hat{T}$.

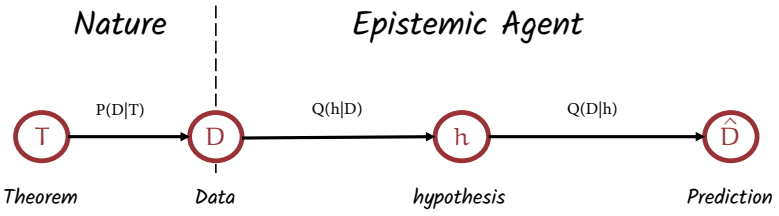


FIGURE 6.2: A hypothesis is the encoded understanding of a *law of nature*.

Moreover, h is falsifiable, as any agent can use h to predict $\hat{D} := \mathrm{Q}(D|h)$. The $\mathrm{Q}(h,\mathrm{D})$ distribution contains the **understanding** of the epistemic agent of the "Law of Nature" ($\mathrm{Q}(h|D)$) and the **expression** of this understanding (the prediction $\mathrm{Q}(D|h)$). In other words, $\mathrm{Q}(D,H)$ defines an encoder (understanding) - decoder (expression).[7]

If other epistemic agents have competing hypothesis $(h_j, h_k, \ldots)$, how should we select the best hypothesis?

The best hypothesis is the one that on average describes D with $\mathrm{H}[D]$ bits. Any hypothesis that take less bits than $\mathrm{H}[D]$ cannot perfectly reconstruct D (underfitting). Any hypothesis that uses more

[5] A lossy encoder or noisy channel are in practice the same.

[6] lossless $\to$ $\hat{T} = X^{-1}(X(\hat{T}))$. Therefore, by RI, $\mathrm{H}[X(\hat{T})] = \mathrm{H}[\hat{T}]$ and for all practical purposes $h := \hat{T}$.

[7] In Machine Learning, the *understanding* happens during training and the *expression* in test time.

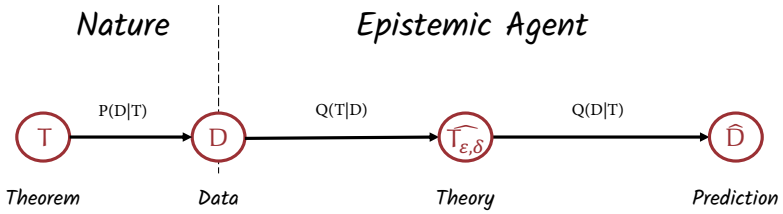


FIGURE 6.3: A theory is a tested hypothesis that predicts the *law of nature* within a margin of error and a level of confidence.

than $H[D]$ bits is adding spurious correlations to the data (overfitting) and might not generalise well.

Besides selecting the best model among the available competitors, the epistemic agent wants to transform her winning hypothesis into a theory that works within a tolerance of error (ϵ) and margin of confidence (δ):

$$\Pr\left[\mathbb{1}_{\hat{T_{\epsilon,\delta}}(D)\neq T(D)} \leqslant \epsilon\right] \geqslant (1-\delta). \tag{6.3}$$

In reality, unfortunately, she can access only a sample S^n of the true distribution of the data $P(D|T)$. How confident can agents be in the performance of h in future data if they can only access the error of h in the sample (past) data?

6.2 PAC-SHANNON

This section will use Shannon's theorems to give PAC bounds to the information-theoretical learning setting presented in the previous section.

We recognise that:

1. ITML setting is equivalent to MDL (which will be described in Section 6.5);
2. using information in the weihgts as a measure of complexity was already discussed by other authors ([Tis20; Ach19; SST10]); and also that
3. Shamir et al. has presented the first PAC formulation of IBT [SST10].

Yet, to the extent of our knowledge, the specific PAC formulation we are about to describe is an original contribution of this dissertation.[8]

Recall Theorem 5.1 (Shannon's 1st Law):

Theorem 5.1 (Shannon's 1st Law). *The optimal binary encoding* $X^k = (X_1, \cdots, X_k)$, $X_i \in \{0,1\}$, *of a n-symbols message* $S^n = (S_1, \cdots, S_n)$, *where*

[Tis20] Tishby, *The Information Bottleneck View of Deep Learning: Why do we need it?*. URL: https://youtu.be/utvIaZ6wYuw

[Ach19] Achille, 'Emergent Properties of Deep Neural Networks'. URL: https://escholarship.org/uc/item/8gb8x6w9

[SST10] Shamir et al., 'Learning and generalization with the information bottleneck'.

[8] Therefore, we took the liberty of naming it PAC-Shannon.

$S_i \in \mathbb{A}_S$ are i.i.d. $\sim p(s)$ has an expected size $k \approx n\mathrm{H}[S]$ for sufficiently large n.

We can rewrite it as:

Theorem 6.1 (1st Shannon PAC formulation)**.** *Let* $\mathrm{X} = x \sim \mathrm{P}(\mathrm{X})$ *and* $\mathrm{S}^n = \{x_1, .., x_n\}, x \sim \mathrm{P}(\mathrm{X})$,

$$Pr\left[\left(\frac{\mathrm{H}[S^n]}{n} - \mathrm{H}[X]\right) > \epsilon\right] < \delta \tag{6.4}$$

The same for Theorem 5.5 (Shannon's 2nd Law):

Theorem 5.5 (Shannon's 2nd Law)**.** *All rates below capacity C are achievable. Specifically, for every rate* $\mathrm{R} < \mathrm{C}$*, there exists a sequence of* $(2^{n\mathrm{R}}, n)$ *codes with maximum probability of error* $\lambda^{(n)} \to 0$*. Conversely, any sequence of* $(2^{n\mathrm{R}}, n)$ *codes with* $\lambda^{(n)} \to 0$ *must have* $\mathrm{R} \leqslant \mathrm{C}$*.*

We can rewrite it as:

Theorem 6.2 (2nd Shannon PAC formulation)**.** *Let two discrete random variables (from finite spaces) represent X, the input, and Y, the output of a stochastic mapping* $\mathrm{Q}(X, Y)$ *(lossy encoder/channel).* $(X, Y) \sim \mathrm{P}(X, Y)$*,* P *is unknown and let* Я[9] *represent the information rate of* Q *(the expected number of bits it needs to represent a symbol of the alphabet* $\mathbb{A}_X$*). Let* $\epsilon(\mathrm{Q})$ *represent the bit error of* Q*. The expected error (the risk) of* Q *under the distribution* $\mathrm{P}(X, Y)$ *is* $\mathbb{E}_{\mathrm{P}(X,Y)}[\epsilon(\mathrm{Q}(X, Y))] = \mathrm{R}(\mathrm{Q})$*:*

$$\forall Я : Я \leqslant \mathrm{I}[X; Y], \tag{6.5}$$

$$\exists \mathrm{Q} : \mathrm{R}_{\mathrm{P}(X,Y)}(\mathrm{Q}(X, Y)) < \epsilon \tag{6.6}$$

[9] We use the symbol Я for rate here to differentiate from the risk R symbol, which is already widespread in the MLT community.

6.2.1 *Shannon guarantees*

$$D \xrightarrow{\mathrm{Q}(W|D)} \underset{\text{hypothesis}}{W} \xrightarrow{\mathrm{Q}(D|W)} \hat{D}$$

$$(X, Y) \longrightarrow W \longrightarrow (X, \hat{Y})$$

Let $D \sim \mathrm{P}(D)$ represent observable data and S^n a sample of n observations from D. We assume that the hypothesis h is parametrised by W and $h(\cdot)$ is a deterministic and invertible function. Let a learning algorithm $\mathcal{A} : D \to W, \mathcal{A} := \mathrm{Q}(W, D)$ generate the hypothesis via W. By reparametrisation invariance (RI), $\mathrm{H}[h(W)] = \mathrm{H}[W]$. If h trained with the sample S^n achieves training error $\epsilon_{\mathrm{S}^n}(h)$. What is the expected out-of-sample error of h, $\epsilon_D(h)$?

For simplification sake, let us assume the supervised case where $D = (X, Y)$ and $Y \in \{0, 1\}$.[10] By Theorem 6.2 (2nd Shannon PAC formulation),

[10] Similar to MLT problem setting.

$$\exists h_{\text{Shannon}} := \mathrm{Q}(\mathrm{D}|w^*) : Я(h_{\text{Shannon}}) = \mathrm{I}[W; D], \text{ and} \tag{6.7}$$

$$\epsilon_D(h_{\text{Shannon}}) = \mathrm{R}_D(h_{\text{Shannon}}) < \epsilon \tag{6.8}$$

Let us remember that $h_{\text{Shannon}} \in \mathcal{H}_{\text{Q}}$ is the theoretical optimal $\text{Q}^*(X|W) \to \hat{Y}$. Also, in Section 5.7, we saw that the channel capacity $\text{I}[W;D]$ defines the number of non-confusable inputs/mappings (or the number of confused mappings limited to a certain ϵ margin), $2^{\text{I}[W;D]}$ Eq. (5.71). We then call $\mathcal{H}_{\text{Q}}^{\delta}$ the typical hypothesis space of h and we know that its cardinality can be computed:

$$\text{Я}^* = \text{Я}(\text{Q}^*) = \text{I}[W;D] \tag{6.9}$$

$$|\mathbb{A}_{\text{Q}}^{\epsilon}| = |\mathbb{A}_Y|^{2^{\text{I}[W;D]}} = 2^{2^{\text{I}[W;D]}} \tag{6.10}$$

Serendipitously, we purposely did not restrict our hypothesis, but Eq. (6.10) produced for us the cardinality of the hypothesis space of the solutions within a tolerance error and confidence. This conclusion is remarkable.

Using Theorem 4.3 ([Hau88], finite space, inconsistent case), we can already give an upper bound to the out-of-sample error of h:

$$\epsilon(h) \leqslant \sqrt{\frac{\ln 2^{2^{\text{I}[W;D]}} + \ln 2/\delta}{2n}} \tag{6.11}$$

$$\approx \sqrt{\frac{2^{\text{I}[W;D]} + const. + \ln 2/\delta}{2n}} \tag{6.12}$$

$$\therefore n \leqslant \frac{2^{\text{I}[W;D]} + \ln 2/\delta}{2\epsilon^2} \tag{6.13}$$

We get a non-vacuous bound as long as $\mathcal{O}(\text{I}[W;D]) < \mathcal{O}(\log n)$. Unfortunately, we cannot access the true $\text{I}[W;D]$; we only access its empirical approximation $\tilde{\text{I}}[w;D]$.

Let us see if we can at least bound the true mutual information between W and D. Let us choose the D_{KL} as a loss function.[11]

[11] For reasons that will be clear in the next chapters.

$$\text{R}_{\text{D}}(h) = \mathbb{E}(\varepsilon(h)) \tag{6.14}$$

$$= \mathbb{E}_{W,D}\left[\text{D}_{\text{KL}}(\text{P}(W,D) \| \text{Q}(W,D))\right] \tag{6.15}$$

$$= \mathbb{E}_{W,D}\left[\log \frac{\text{P}(W,D)}{\text{Q}(W,D)}\right] \tag{6.16}$$

$$= -\mathbb{E}_{W,D}\left[\log \frac{\text{Q}(W,D)}{\text{P}(W,D)}\right] \tag{6.17}$$

$$= -\mathbb{E}_{W,D}\left[\log \text{Q}(W,D) - \cancelto{\text{const.}}{\log \text{P}(W,D)}\right] \tag{6.18}$$

$$= -\mathbb{E}_{W,D}\left[\log \text{Q}(W,D)\right] = \text{I}[W;D] \tag{6.19}$$

$$\text{R}_{S^n} = -\mathbb{E}_{W,S^n}\left[\log \text{Q}(W,S^n)\right] = \tilde{\text{I}}[w;D] \tag{6.20}$$

$$\mathcal{O}\left(|\text{R}_{S^n}(h) - \text{R}_{\text{D}}(h)|\right) = \mathcal{O}\left(|\tilde{\text{I}}[w;D] - \text{I}[W;D]|\right) \tag{6.21}$$

From Theorem 6.1 (1st Shannon PAC formulation),

$$\Pr(|\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]| > \epsilon) < \delta \tag{6.22}$$

Now we can follow [Hau88] steps as we did in Section 4.7.

[Hau88] Haussler, 'Quantifying inductive bias: AI learning algorithms and Valiant's learning framework'.

Theorem 6.3 (PAC Shannon, finite space, consistent case)**.** *Let $\mathcal{A}$ be a learning algorithm that returns a consistent hypothesis h, i.e. $\hat{\mathrm{R}}_{\mathrm{S}}(h) = 0$, for any hypothesis h and unknown distribution $\mathrm{D} = \mathrm{P}(X, Y)$. Let $|\mathrm{S}| = n$, then, $\forall n \geqslant \mathrm{N}_0$:*

$$Pr[h \in \mathcal{H} : \mathrm{R}_{\mathrm{D}}(h) > \epsilon] < e^{-\epsilon n + \tilde{\mathrm{I}}[w;D]} \tag{6.23}$$

Proof. Let h be parametrised by W and the empirical mutual information of the weights w.r.t the available sample S^n be $\tilde{\mathrm{I}}[w;D]$. From Eq. (6.22), let us call h_{bad} a consistent hypothesis that does not generalises and $\mathcal{H}_{\text{bad}}$ the space of all possible bad hypotheses.

$$Pr\left(|\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]| > \epsilon\right) < \delta \therefore \tag{6.24}$$

$$Pr\left[\mathrm{R}(h) = 0 \wedge |\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]| > \epsilon\right] = 1 - \delta \tag{6.25}$$

$$\mathbb{E}_{\mathrm{S}}[\mathrm{R}(h) = 0 \wedge |\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]| > \epsilon] = (1-\delta)^n \tag{6.26}$$

$$\mathbb{E}_{\mathrm{D}}[\mathrm{R}(h) = 0 \wedge |\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]| > \epsilon] = |\mathcal{H}_{\text{bad}}|(1-\delta)^n \tag{6.27}$$

$$\tag{6.28}$$

Fortunately, we know how to find the cardinality of $\mathcal{H}_{\text{bad}}$. $\tilde{\mathrm{I}}[w;D]$ is our channel capacity, *i.e.* the number of typical different encodings (or transformations) we can have. Every transformation of an input X can lead to $|\mathbb{A}_Y|$ values. Therefore, $|\mathcal{H}_D^\delta| \approx 2^{\mathrm{I}[W;D]}$ and $|\mathcal{H}_{\mathrm{S}^n}^\delta| \approx 2^{\tilde{\mathrm{I}}[X;Y]}$. Consequently, $|\mathcal{H}_{\text{bad}}| = 2^{|\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]|}$. From where we follow:

$$\mathbb{E}_{\mathrm{D}}[\mathrm{R}(h) = 0 \wedge |\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]| > \epsilon] = |\mathcal{H}_{\text{bad}}|(1-\delta)^n \tag{6.29}$$

$$= |\mathcal{H}_{\text{bad}}|\, e^{-\epsilon n} \tag{6.30}$$

$$\epsilon < e^{-\epsilon n + 2^{|\tilde{\mathrm{I}}[w;D] - \mathrm{I}[W;D]|}} \tag{6.31}$$

As we already said, $\mathrm{I}[W;D]$ is intractable, but we still can get a bound:

$$\epsilon < e^{-\epsilon n + 2^{\tilde{\mathrm{I}}[w;D]}} \tag{6.32}$$

□

Theorem 6.4 (PAC Shannon, finite space, consistent case: sample complexity)**.** *A learning algorithm $\mathcal{A}$ can learn task with:*

$$n < \frac{1}{\epsilon}\left(\tilde{\mathrm{I}}[w;D] + \ln\frac{1}{\delta}\right)$$

training examples.

Proof.

$$\delta > e^{-\epsilon n + 2^{\tilde{I}[w;D]}} \tag{6.33}$$

$$\ln \delta > -\epsilon n + 2^{\tilde{I}[w;D]} \tag{6.34}$$

$$\epsilon n < 2^{\tilde{I}[w;D]} - \ln \delta \tag{6.35}$$

$$n < \frac{1}{\epsilon}(2^{\tilde{I}[w;D]} - \ln \delta) \tag{6.36}$$

$$n \in \mathcal{O}\left(\frac{1}{\epsilon}\left(2^{\tilde{I}[w;D]} - \ln \delta\right)\right) \tag{6.37}$$

□

Theorem 6.5 (PAC Shannon, finite space, inconsistent case)**.** *Let $\mathcal{A}$ be a learning algorithm that returns an inconsistent hypothesis h, i.e. $\hat{R}_S(h) > 0$, for any hypothesis h and unknown distribution $D = P(X, Y)$. Let $|S| = n$, then, $\forall n \geqslant N_0$:*

$$\epsilon < \sqrt{\frac{2^{\tilde{I}[w;D]} + \ln \frac{2}{\delta}}{2n}} \tag{6.38}$$

Proof. Using the Chernoff-Hoeffding inequality and the union bound as Theorem 4.1 and Section 4.7, we have:

$$Pr[w \in W : |\tilde{I}[w;D] - I[W;D]| > \epsilon] < 2e^{-2n\epsilon^2} \tag{6.39}$$

$$Pr[w \in W : |\tilde{I}[w;D] - I[W;D]| > \epsilon] = \frac{1}{|\mathbb{A}^{\epsilon}_{\mathcal{H}}|}\delta \tag{6.40}$$

$$2e^{-2n\epsilon^2} < |\mathbb{A}^{\epsilon}_{\mathcal{H}}|^{-1}\delta \tag{6.41}$$

$$ln2 - 2n\epsilon^2 < -\ln |\mathbb{A}^{\epsilon}_{\mathcal{H}}| + \ln \delta \tag{6.42}$$

$$\epsilon < \sqrt{\frac{\ln |\mathbb{A}^{\epsilon}_{\mathcal{H}}| + \ln \frac{2}{\delta}}{2n}} \tag{6.43}$$

$$\epsilon < \sqrt{\frac{\ln 2^{2^{\tilde{I}[w;D]}} + \ln \frac{2}{\delta}}{2n}} \tag{6.44}$$

$$\epsilon < \sqrt{\frac{2^{\tilde{I}[w;D]} \cancel{\ln 2}^{1} + \ln \frac{2}{\delta}}{2n}} \tag{6.45}$$

□

Theorem 6.6 (PAC Shannon, finite space, inconsistent case: sample complexity)**.** *A learning algorithm $\mathcal{A}$ can learn task with:*

$$n < \frac{2^{\tilde{I}[w;D]} + \ln \frac{2}{\delta}}{2\epsilon^2} \tag{6.46}$$

training examples.

Proof.

$$\epsilon^2 < \frac{2^{\tilde{I}[w;D]} + \ln \frac{2}{\delta}}{2n} \tag{6.47}$$

$$n < \frac{2^{\tilde{I}[w;D]} + \ln \frac{2}{\delta}}{2\epsilon^2} \tag{6.48}$$

□

6.3 "REALS" ARE NOT REALLY A PROBLEM

A possible weakness of the proposed ITML perspective is that we limited the space of the data D to a finite set (discrete random variable).

Foremost, there is a mathematical argument [Cha06, pp. 99–116] against the physical existence of a "continuum": after all, some real numbers are uncomputable [Tur36].[12] Similarly, in Section 6.1, we argued that there was no pointing in learning a concept that could not fit the finite epistemic agents' minds.

MLT, however, is agnostic to the unknown distribution, hence, it can be a continuous function. Bayes' rule is the same for probability mass functions (pmfs) and probability density functions (pdfs) after all [Mac02; Valoo]. However, when models use continuous random variables, there is no sense in choosing "the most probable model": the probability of a continuous random variable tends to zero at any single point. Only a nonzero range has a nonzero probability. As [Valoo] puts it: "(⋯) a high density *per se* is not important, but the overall probability mass in the vicinity of a model is."

Rissanen gave a more formal version of this justification. He noticed that we could always choose a (n, k) code such that the quantisation error of the real distribution is within a margin of error. Imagine the dataset $S^{(n)}$ is sampled from a continuous distribution $D = f(x)$ and there is a uniform distribution encoder (raw bit encoder) $U(D)$ that encodes D into a code of k bits.

$$U(x) = \frac{f(x)}{2^k} \tag{6.49}$$

$$\mathrm{H}[\mathrm{U}(S^{(n)})] = -\sum_{x_1}^{x_n} \log \frac{2^k}{f(x)} = -n(\log f(x) - \log 2^k) \tag{6.50}$$

$$Pr\left[\left(\frac{\mathrm{H}[\mathrm{U}(S^n)]}{n} - \mathrm{H}[\mathrm{U}(x)]\right) \leqslant \log \frac{1}{2^k} > \epsilon\right] < \delta \tag{6.51}$$

This is Shannon's argument that for sufficiently large n and we can always *digitise* the sample to a desired small tolerance of error ϵ, Theorem 6.1 (1st Shannon PAC formulation).

[Cha06] Chaitin, *Meta Math! The Quest for Omega*.

[Tur36] Turing, 'On Computable Numbers, with an Application to the Entscheidungsproblem'.

[12] Another remark is that no physical quantity has ever been measured with more than twenty digits of precision [Cha06, p. 92].

[Mac02] MacKay, *Information Theory, Inference, and Learning Algorithms*.

[Valoo] Valpola, 'Bayesian Ensemble Learning for Nonlinear Factor Analysis'.

6.4 INFORMATION MEASURES THE COMPLEXITY OF TASKS

In Section 6.2, we proved that information measures the complexity of a task. The information-complexity relation, however, was already presented in [Ris86; HVC93], and goes back to [WB68] (according to [Maco2; Valoo]).

In our setting, Nature is a "supervisor" who knows the true distribution of the data $P(D)$ and send us a message D (the observations). The message D implicitly carries the intrinsic pattern $P(D)$ that governs it. Our epistemic agent comes up with an hypothesis h_i that predicts observations $Q(D|h_i)$.

[Ris86] Rissanen, 'Stochastic complexity and modeling'.

[HVC93] Hinton and Van Camp, 'Keeping the neural networks simple by minimizing the description length of the weights'.

[WB68] Wallace and Boulton, 'An Information Measure for Classification'.

6.4.1 *Minimal Description Length Principle*

Supose there is a supervisor (sender) who wants to transmit a given data (D) to a receiver. The supervisor will use a model to compress the data, but will also need to send the misfit bits of the model prediction to the data.

The Minimum Description Length Principle [Ris86] asserts that the best model for a data distribution minimises the combined cost of describing the model and describing the misfit between the model and the data.[13] $P_\theta(D) = P(D|\theta)$ determines the probability of the observation D. Imagine there a statistical model of the real P_θ parametrised by w, $P(w|\theta)$. The supervisor send a message with:

[13] Now we give the proper attribution to this idea already presented in Sections 6.1 and 6.2.

1. $L(\theta)$ bits pertaining which model $h(w)$ to use;
2. $L(D|\theta)$ bits corresponding to the data D predicted by the model, which can be further subdivided onto:
   a) Parameter block: $L(w|\theta) = -\log P(w|\theta)\delta w$;
   b) Data misfit block: $L(D|w,\theta) = -\log P(D|w,\theta)\delta D$.

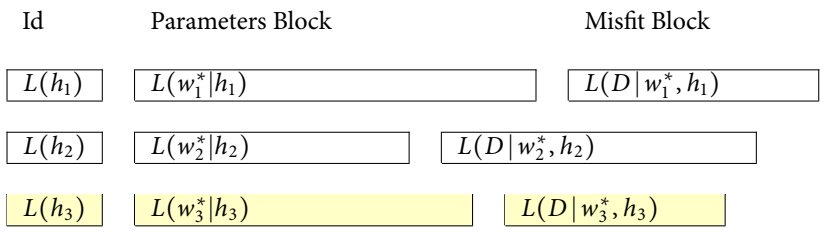


FIGURE 6.4: Comparing hypotheses with the minimum description length principle. Adapted from [Maco2].

There is a clear tradeoff between the parameter block and the data misfit (see Figure 6.4): models with fewer parameters (large

δw) have smaller parameter blocks but do not fit the data as well and therefore have larger misfit blocks; conversely, over parametrised models (small δw) have larger parameter blocks, but smaller misfit blocks. The optimal description minimises the combined length of the parameter and data misfit blocks (Figure 6.4, h_3).

Correspondence to Bayesian inference

Thus, Rissanen's complexity is[14] $\mathrm{L}(D, \theta) = \mathrm{L}(\theta) + \mathrm{L}(D|\theta)$. In a Bayesian interpretation, the length $\mathrm{L}(\theta)$ for different h defines an implicit prior $\mathrm{P}(\theta)$ over alternative hypotheses [Mac02]. If there is no bias towards one or another hypothesis, $\mathrm{P}(\theta) = 2^{-\mathrm{L}(\theta)}$ is uniform and the identifier for the model has the same "cost" $\mathrm{L}(\theta)$. Likewise, $\mathrm{L}(D|\theta)$ defines the density $\mathrm{P}(D|\theta)$ that relates to the evidence for each hypothesis.

In other words, message lengths can be mapped onto posterior probabilities:

$$\mathrm{L}(D, \theta) = -\log \mathrm{P}(\theta) - \log(\mathrm{P}(D|\theta)\delta D) \tag{6.52}$$

$$= -\log \mathrm{P}(D|\theta) + \text{const.} \tag{6.53}$$

As a consequence, MDL has always a Bayesian model comparison interpretation, and *vice-versa*.

6.5 MINIMUM DESCRIPTION LENGTH LEARNING

Using the MDL principle, [HVC93] proposed an information-theoretical machine learning framework.

Notice that in the MDL coding scheme (Section 6.4.1), to send the value of δw which is *arbitrarily* small, we will need an encoding that can lead to *arbitrarily* long messages.

The bits-back argument

To avoid this potential peril, Hinton and Van Camp propose the following coding scheme where a decodable message is obtained without encoding δw:

1. The sender computes a distribution $\mathrm{Q}(W|D, \theta)$ based on observations of D.[15]
2. The sender draws a random sample w from $\mathrm{Q}(W|D, \theta)$ and encode it with $\mathrm{P}(w|\theta)$.
3. The sender encodes D using $\mathrm{P}(\mathrm{D}|w, \theta)$.

[14] Also known as Stochastic Complexity.

[Mac02] MacKay, *Information Theory, Inference, and Learning Algorithms*.

[HVC93] Hinton and Van Camp, 'Keeping the neural networks simple by minimizing the description length of the weights'.

[15] We will explain how to compute this distribution later.

The trick is that in the second step, instead of using random bits to choose w from $\mathrm{Q}(W|D,\theta)$, the sender can use a secondary message as the random bits. So, a long communication, we can say that on average the cost (or length) of the messages are:

$$\mathrm{L}(w|\theta) \quad +\mathrm{L}(\mathrm{D}|w,\theta) - \text{'bits back'} \tag{6.54}$$

$$\mathrm{L}(w|\theta) \quad +\mathrm{L}(\mathrm{D}|w,\theta) - \mathrm{L}(w|\mathrm{D},\theta) \tag{6.55}$$

$$-\log \mathrm{P}(w|\theta)\delta w \quad -\log \mathrm{P}(\mathrm{D}|w,\theta)\delta D - \log \mathrm{P}(w|\mathrm{D},\theta)\delta w \tag{6.56}$$

$$-\frac{\log \mathrm{P}(w|\theta)\cancel{\delta w}\,\mathrm{P}(\mathrm{D}|w,\theta)\delta D}{\mathrm{P}(w|\mathrm{D},\theta)\cancel{\delta w}} \tag{6.57}$$

$$-\log \mathrm{P}(\mathrm{D}|\theta)\delta D \tag{6.58}$$

$$-\log \mathrm{P}(\mathrm{D}|\theta) \cancel{-\log \delta D} \xrightarrow{const.} \tag{6.59}$$

$$\tag{6.60}$$

Thus the proposed coding scheme yields the optimal description length. The only missing step is how the sender computes the distribution Q.

For that, [HVC93] proposes using the Kullback-Leibler divergence (D_{KL}) as a loss function:

$$\ell = \mathrm{D}_{\mathrm{KL}}(\mathrm{Q}||\mathrm{P}) \tag{6.61}$$

to approximate the parametric Q to the real P. This method for parametric approximation of posterior pdfs was called *ensemble learning* and is more commonly known as *variational learning*.

6.5.1 *Shannon, Kolmogorov-Chaitin and Rissanen complexities*

Let us remind ourselves that Shannon's information measures the expected number of bits needed for encoding a random variable D, *i.e.* the entropy $\mathrm{H}_p(D)$ is the expected length of D in bits using the optimal encoder p.

From Eq. (5.55):

$$2^{-n(\mathrm{H}[D]+\epsilon)} \leqslant \mathrm{P}\left(S^{(n)}\right) \leqslant 2^{-n(\mathrm{H}[D]-\epsilon)} \tag{6.62}$$

$$2^{-\mathrm{H}[D]} < 2^{-(\mathrm{H}[D]+\epsilon/n)} \leqslant \mathrm{P}\left(D\right) \leqslant 2^{-(\mathrm{H}[D]-\epsilon/n)} < 2^{-\mathrm{H}[D]+1} \tag{6.63}$$

$$2^{-\mathrm{L}^*(D)} \leqslant \mathrm{P}\left(D\right) \leqslant 2^{-\mathrm{L}^*(D)+1} \tag{6.64}$$

However, one can use a non-optimal encoder q for which the expected length is $\mathrm{H}_{p,q}(D)$. Each encoder/decoder q can be seen as a

"program" that ouputs an average number of bits $L(D|q) = L_q(D) = H_{p,q}(D)$. The minimum program that outputs D or *minimum description length* of D is $L^*(D) = L_p(D) = H_{p,p}(D) = H_p(D)$.

In Section 5.8.1, we mentioned the algorithmic information perspective where Kolmogorov-Chaitin complexity (KC) measures the length of the shortest computer program P which is capable of producing the data D. Therefore,

$$P(D) = 2^{-KC(D)} \tag{6.65}$$

A well-known algorithmic information result is that KC is not computable due to the halting problem [Tur36; Cha06]. Therefore, we cannot know if a learning algorithm that halts when finding the best $P(D) = 2^{-KC(D)}$ will ever halt. This relates to the fact that the Shannon information needed to describe a continuous random variable is infinite.

Confirming Mitchel's theorem [Mit80], a bias on P is needed. Either $P(D)$ is binned into a probability mass function (therefore, biased by its precision δD), or $P(D)$ is a statistical model, *i.e.* it is a "family" of functions identified by a parameter vector θ, $P(D|\theta)$.[16] The first case leads to Shannon Information as a complexity measure (where the prediction should ensemble all encoder/decoders q_i weighted by their posterior probabilities $P(D|q_i)$). The second case, to the idea of *stochastic complexity* developed by Rissanen [Ris86] (where instead of averaging over all possible programs, the prediction assumes the best encoder/decoder $P(D) = P(D|q^*)$).

Shannon's entropy, Kolmogorov-Chaitin's complexity, and Rissanen's Stochastic complexity are different but related task complexity measures.

[Tur36] Turing, 'On Computable Numbers, with an Application to the Entscheidungsproblem'.

[Cha06] Chaitin, *Meta Math! The Quest for Omega*.

[Mit80] Mitchell, *The Need for Biases in Learning Generalizations*.

[16] In this our conversation, θ was the truth T.

[Ris86] Rissanen, 'Stochastic complexity and modeling'.

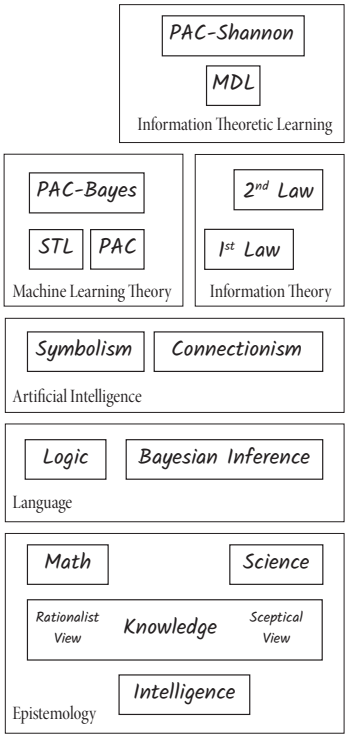


FIGURE 6.5: ITML applies IT to explain machine learning.

[Valoo] Valpola, 'Bayesian Ensemble Learning for Nonlinear Factor Analysis'.

6.6 CONCLUDING REMARKS

This chapter presented the information-theoretical perspective of learning and provided a bridge of this perspective to Machine Learning Theory (MLT).

The previous chapter (Chapter 5) had already shown that information is a measure of change in belief which is also the description length of the data (using the expected negative logarithm of its distribution); therefore, a measure of the data structure or lack of pattern. Any learning method derived from IT can be translated to a Bayesian interpretation by a change of scale [Valoo]. Prior probabilities translate to a coding scheme that is needed to "decode" the data. In other words, information is a measure of complexity of a task. We related

this Shannon Information complexity to Kolmogorov-Chaitin complexity and Rissanen's Stochastic complexity.[17]

In the context where learning is a conversation with Nature (Section 6.1), we used Shannon's theorems to demonstrate that information measures the complexity of the task. MLT and ITML are *two sides of the same coin*. If in MLT we make no assumptions on the task and depend on the hypothesis space, ITML does not assume any hypothesis space but is task-dependent. Either way, learning is about finding patterns in data, and the best hypothesis to describe the data regularities is also the one that compresses it the most.

We presented the MDL framework which was the first Information-Theoretical Machine Learning (ITML) proposed method. And showed from the correspondence of MDL with Bayesian inference.[18]

Therefore, even before introducing IBT, we can conclude that anything that is explainable by it can be explained in current MLT. If so, *what is the purpose of IBT?* After all, according to [Mac02], MDL "has no apparent advantage" beyond as a "pedagogical tool". Why would IBT be any different?

The purpose of IBT (and MDL) is to bring a new narrative. Take a look at the transition from Figure 6.1 to Figure 6.2. If two hypotheses generate the same result, do they represent the same understanding? In practice, yes, they do and we can address them mathematically. [19] But if we think of understanding as meaning, not necessarily.

This other "philosophical" interpretation is understandingly not addressed by the literature. We will, nevertheless, indulge ourselves with some digression. Take, for example, the Lorenz' Ether Theory (LET) and Einstein's Special Relativity Theory (SR).[20] There is simply no way of distinguishing LET or SR experimentally, but there is a philosophical distinction between the two [Sza11] (as cited by [Dal]). In this example, we can return the same question: *What is the purpose of Special Relativity Theory?*

Meanings are not part of the truth we find in Nature but represent the ideally noiseless encoding of our understanding that we create for other epistemic agents to decode. In this sense, just as the sweetness in honey (Section 2.2.3), meaning is projected. It is improbable that the decoded understanding in two "epistemic minds" are the same and different narratives are capable of sparking different analogies and connections.

[17] We will also show that Fisher information is the stochastic complexity for isotropic Gaussian distributions (Section 8.6).

[18] This relation was expected since we already had shown the correspondence of IT and Bayesian inference.

[Mac02] MacKay, *Information Theory, Inference, and Learning Algorithms*.

[19] Remember that Shannon decided not to address meaning in his theory.

[20] Lorentz ether theory describes a universe in which light moves through a medium called ether. The problem is that the ether can be seen as a mathematical construct that can not be measured or observed. It is used to facilitate predictions calculations. Those predictions in the movement of light can be measured. Einstein's Special Relativity describes a new geometry of a universe that has no ether. However, it uses Lorentz mathematical construct to do so.

[Sza11] Szabó, 'Lorentzian Theories vs. Einsteinian Special Relativity — A Logico-empiricist Reconstruction'.

[Dal] Dale, *Are Lorentz aether theory and special relativity fully equivalent?*.
URL: https://physics.stackexchange.com/q/525808

Part III

THE EMERGENCE OF A THEORY

7

The Information Bottleneck Principle

As we already discussed (Section 5.2.1), Shannon intentionally left out from information theory[1] issues of meaning or relevance, and focused on the problem of transmitting information.

Contrarily, Tishby et al. argue in [TPB99] that lossy source compression provides a natural quantitative approach to the matter of relevance and, therefore, they use Information Theory itself to address relevance.

This chapter will present the Information Bottleneck Principle, the foundation of the emergent theory subject of this dissertation. The IB principle approach is related to Rate-Distortion Theory (RDT). Hence, first we will briefly overview RDT as Tishby et al. describe it [TPB99; Slo02]. Then, we will formally present the IB Principle, its problem setting and analytical solution, and show how it can be seen as a particular case of Rate-Distortion Theory.

[1] Which Shannon has always referenced as Communication Theory.

[TPB99] Tishby et al., 'The Information Bottleneck Method'.

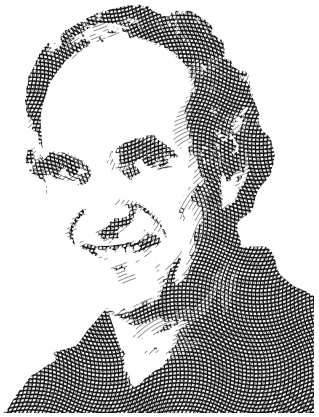


Figure 7.1: Naftali Tishby.

[Slo02] Slonim, 'The information bottleneck: Theory and applications'.

7.1 Rate-Distortion Theory: Relevance through a Distortion Function

We know from Eq. (5.97) that for any rate $R \leqslant H[X]$ there will be a loss in the reconstructed signal. Rate-Distortion Theory (RDT) addresses the problem of determining the rate R that should be communicated

over a channel so that the source (input signal X) can be approximately reconstructed without exceeding an expected distortion.

7.1.1 *The Rate-Distortion Theory (RDT) problem*

Problem setting

1. Let the discrete random variable X denote the **source** of vectors randomly drawn from a probability distribution $p(x)$;
2. Each vector $x \sim p(x)$ is a **message** (signal) you want to transmit among a set of possible messages $\mathbb{A}_X$, *i.e.* $x \in \mathbb{A}_X$;
3. Let another discrete random variable Z denote[2] a compressed **representation** of X ;
4. This representation is defined by a **channel** $p(z|x)$, a stochastic mapping between each message $x \in \mathbb{A}_X$ to each code $z \in \mathbb{A}_Z$;
5. The **rate** R is the channel capacity, i.e. the average number of bits per element $x \in \mathbb{A}_X$ needed to specify a compressed element (code) $z \in \mathbb{A}_Z$.
6. Let $d : \mathbb{A}_X \times \mathbb{A}_Z \to \mathbb{R}^+$ be a function that denotes the **distortion measure** between X and its representation Z. Examples of distortion measures are the mean square error, $d_{\text{MSE}}(x;z) = \langle (x-z)^2 \rangle$ or the Hamming distortion (probability of error) $d_{\text{H}}(x,z) = \mathbb{1}_{[x \neq z]}$.

[2] This compressed representation of X is usually denoted by Z, T or $\hat{X}$

$$X \xrightarrow{\text{channel}} Z$$

Problem Statement

Given the problem setting above, the RDT problem[3] is to find the minimal number of bits per symbol (rate R) that should be communicated over a channel so that the source X can be approximately reconstructed via a representation Z without exceeding an expected distortion D, defined by the distortion function $d(x;z)$.

[3] First defined by Shannon [Sha48].

7.1.2 *Understanding the RDT problem*

The core of the RDT problem is the need for a good compressed representation of a message. From Eq. (5.97), any rate $\text{I}[Z;X] \leqslant \text{H}[X]$ will imply a loss in the reconstructed signal, an expected distortion, $\langle d(x;z) \rangle$.

As we have seen in Section 5.7, low values of $\mathrm{I}[Z;X]$, calculated based on the joint distribution $p(x,z) = p(x)p(z|x)$, imply compact representations, i.e. $|\mathbb{A}_Z|$ is small. In the extreme, all messages are translated to the same code: $|\mathbb{A}_Z| = 1$ and $\mathrm{I}[Z;X] = 0$. Contrastingly, high values of $\mathrm{I}[Z;X]$ imply low compression. In the extreme, Z simply copies X: $\mathrm{I}[Z;X] = \mathrm{H}[X]$ and $|\mathbb{A}_Z| = |\mathbb{A}_X|$.

Suppose we can compress the input data to any amount of information from 0 to $\mathrm{H}[X]$. What will define the relevance of information is the additional constraint of the problem: the distortion measure. Given such function, the partitioning of X defined by $p(z|x)$ has the *expected distortion*:

$$\langle d[x;z]\rangle_{p(x,z)} = \sum_{x,z} p(x)p(z|x)d[x;z] \tag{7.1}$$

Consequently, we are assuming that the definition of relevance is part of the problem setting. In other words, RDT is agnostic on any arbitrary choice of the distortion function. This choice, nevertheless, determines the relevant features of the signal[4] and should be somehow related to the task we want to perform with the input. Thus, **an arbitrary distortion function is, in fact, an arbitrary feature selection** [TPB99].

As we will se further (Section 7.2), Tishby et al. [TPB99] propose a way to cope with this potential pitfall.

[4] The same can be said of a learning algorithm loss function in MLT, which determines what is relevant to be learned.

[TPB99] Tishby et al., 'The Information Bottleneck Method'.

7.1.3 *RDT as a variational problem*

Definition 7.1. The **rate-distortion function**, denoted by $\mathrm{R}(\mathrm{D})$ is defined as:

$$\mathrm{R}(\mathrm{D}) \equiv \min_{p(z|x):\ \langle d(x;z)\rangle \leqslant \mathrm{D}} \mathrm{I}[Z;X]. \tag{7.2}$$

Therefore, $\mathrm{R}(\mathrm{D})$ is the minimum achievable rate among all normalised conditional distributions, $p(z|x)$, for which the distortion constraint is satisfied. The *rate-distortion function* is a non-increasing convex function of D in the *distortion-compression plane* [CT06] (see Figure 7.2).[5]

The region above the curve corresponds to all achievable *distortion-compression* pairs, while below the curve is the non-achievable region. Let $\{\mathrm{D}, \mathrm{I}_X\}$ be a *distortion-compression* pair, if it is in the achievable region, there is a representation Z with a compression level $\mathrm{I}[Z;X] = \mathrm{I}_X$ and an expected distortion of at most D. If it is in the non-achievable

[CT06] Cover and Thomas, *Elements of Information Theory*.

[5] We will explain what β means later.

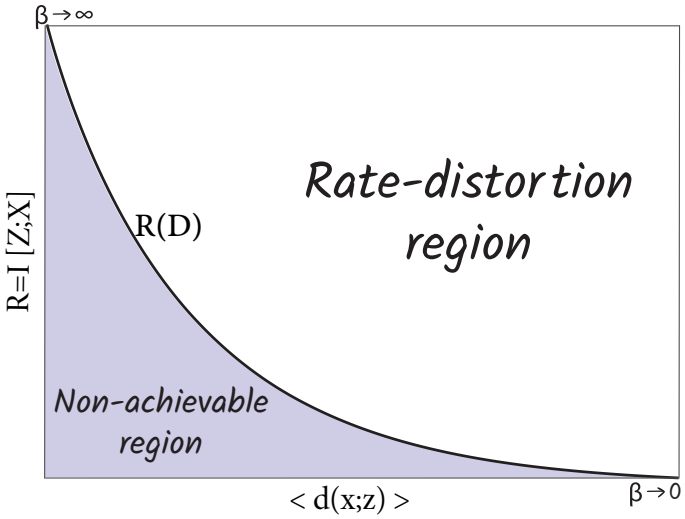


FIGURE 7.2: The rate-distortion function R(D) in the distortion-compression plane.

region, there is no such representation Z. This limit on the achievability of representations is a direct consequence of Shannon's laws (5.2).

Instead of solving the minimisation problem in (7.2) exactly, the problem is usually approximated by the following Lagrangian relaxation functional:

$$\mathcal{F}[p(z|x)] = \mathrm{I}[Z;X] + \beta\langle d(x;z)\rangle_{p(x,z)}, \tag{7.3}$$

under the normalisation constraint $\sum_z p(z|x) = 1, \forall x \in \mathbb{A}_X$.

[TPB99] Tishby et al., 'The Information Bottleneck Method'.

Theorem 7.1. *The solution of the variational problem [TPB99]*

$$\frac{\partial \mathcal{F}}{\partial p(z|x)} = 0, \tag{7.4}$$

for normalised distributions $p(z|x)$ is given by the exponential form

$$p(z|x) = \frac{p(z)}{\mathrm{Z}(x,\beta)} \exp(-\beta\, d(x;z)), \tag{7.5}$$

where Z is the normalisation factor (partition function). The Lagrange multiplier β is positive and

$$\frac{\partial \mathrm{R}}{\partial \mathrm{D}} = -\beta. \tag{7.6}$$

This is an implicit solution[6] as $p(z)$ on the right-hand side of Eq. (7.5) depends on $p(z|x)$[7].

[6] Implicit solution means a solution in which dependent variable is not separated.

[7] $p(z) = \sum_{x,z} p(z|x)\, p(z)$

7.2 THE IB PRINCIPLE: RELEVANCE THROUGH A TARGET VARIABLE

The problem of extracting what is relevant from data depends on a suitable definition of relevance. The main weakness of the RDT ap-

proach is that it addresses relevance through a distortion function that is not related to a specific task at hand.

The **IB Principle**, suggested by Tishby et al. [TPB99] introduces an alternative approach: defining a "target" variable is simpler and more direct than defining a distortion measure.

For example, in speech compression[8], any compression beyond the signal's entropy cannot be perfectly reconstructed; it is a lossy compression. On the other hand, a transcript has orders of magnitude lower entropy than the acoustic waveform, which means that for the task of understanding what has been transmitted, it is possible to compress the signal *much* further without losing any information about meaning [TPB99].

[8] By the time of [TPB99] publication, Tishby was working on speech-related problems.

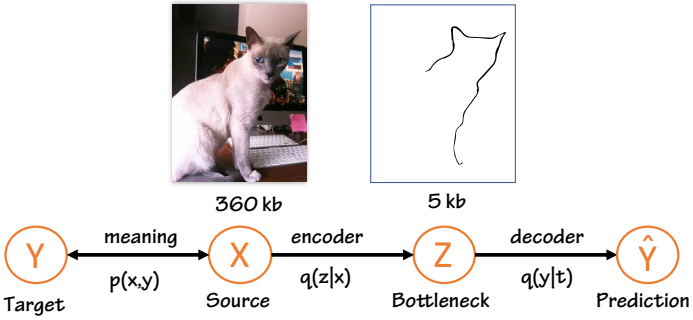


FIGURE 7.3: The IB problem setting.

In many situations, we have access to an additional variable that determines what is relevant. If we want to recognize cats in pictures, maybe we do not need a 360 kb picture as depicted on the left in Figure 7.3; the 5 kb representation on the right may suffice. The exact representation would not be sufficient for the task of recognizing the breed of the cat, in any case. **Relevance is task-dependent**.

7.2.1 *The IB Problem Setting*

Definitions

1. Let X be a random variable that denotes the **Source** [9] of **messages** $x \in \mathbb{A}_X$;
2. Let Y be a random **relevant variable** (or **Target**) that defines the intended meaning $p(y|x)$ of the message x;
3. Let Z be an **information bottleneck** variable, the representation, that obeys the Markov chain $Y \leftrightarrow X \leftrightarrow Z$;
4. Let the conditional p.d.f $p(z|x)$ be the **encoder**, *i.e.* a stochastic mapping from each value of $x \in \mathbb{A}_X$ to a codeword $z \in \mathbb{A}_Z$;

[9] The IB problem is a one-shot coding problem, the operations are performed letterwise [ZEASS20].

5. $\mathrm{I}[Z;X]$ is the **rate** (or compression level) of the encoder, and reflects how much the bottleneck representation Z compresses X;

6. Let the conditional p.d.f $p(y|z)$ be the **decoder**, *i.e.* a stochastic mapping from each value of $z \in \mathbb{A}_Z$ to a prediction $\hat{y} \in \mathbb{A}_Y$;

7. $\mathrm{I}[Z;Y]$ is the **relevant information** that the compressed representation Z keeps from the label variable Y;

Assumptions

i. The random variables X, Y and Z, are **discrete**;

ii. $\mathbb{A}_X$, $\mathbb{A}_Y$ and $\mathbb{A}_Z$ are **finite sets**;

iii. X and Y are dependent, and **the joint distribution** $\mathrm{P}(X = x, Y = y) = p(x, y)$ **is known**;

iv. The source X is an ergodic process[10]; therefore $x \sim p(x)$ are not necessarily mutually independent.

v. The encoder and the decoder are stochastic mappings. Hence, act like noisy channels.[11]

[10] The *ergodic* property means statistical homogeneity [SW49]: its statistical properties can be deduced from a single, sufficiently long, random sample of the process.

[11] Notice that given the Markov chain $Y \leftrightarrow X \leftrightarrow Z$, due to reparemetrisation invariance (Theorem 5.4 (reparametrisation invariance (RI))), a deterministic mapping of the data does not throw out information, i.e. let $f : \mathbb{A}_X \to \mathbb{A}_Y$ be deterministic, $\mathrm{I}[f(X);Y] = \mathrm{I}[X;Y]$.

$$\underset{\text{target}}{Y} \xrightarrow{\text{channel}} \underset{\text{signal}}{X} \xrightarrow{\text{channel}} Z$$

Problem statement

The *information bottleneck problem* consists of finding an encoder $p(z|x)$ that produces a codebook Z that compress X as much as possible, i.e. $\mathrm{I}[Z;X]$ is minimal, while keeping the *relevant information* of X for predicting Y, $\mathrm{I}[Z;Y]$. In other words, the representation Z acts like a **bottleneck** that "squeezes" the relevant information that X contains about the target Y in a compressed form, hence the name "information bottleneck".

7.2.2 *Relation to other Information Theory Problems*

Connections between problems allow extending ideas from one setup to another. In this regard, the IB problem is closely related to other coding problems like the *Indirect* or the *Remote Source-coding problem*, also known as the *CEO Problem*, and the *privacy funnel problem* [ZEASS20].

[ZEASS20] Zaidi et al., 'On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views'.

7.2.3 *Relation to Minimum Sufficient Statistics*

In the IB problem, the target variable is what we want to predict. Y acts as a parameter of X and $Z \perp Y$. Thus, *the representation Z is a statistic of X*.

For Z to be a **sufficient statistic** of X w.r.t. Y, it must preserve all relevant information in X, $\mathrm{I}[\mathrm{Y};\mathrm{X}] = \mathrm{I}[\mathrm{Z};\mathrm{X}]$. In other words, no other statistic of X can provide any additional information as to the value of Y then Z does.

The representation is **minimal** if it is the smallest among all possible representations.

Therefore, we can say that the information bottleneck is the problem of finding the *minimum sufficient statistics* of the random variable X w.r.t Y, and therefore, IB Lagrangian gives the minimum approximately sufficient statistic.

7.3 THE IB CURVE

As in RDT, the compactness of the representation is measured by $\mathrm{I}[Z;X]$. The distortion upper bound constraint, however, is replaced by a lower bound constraint over the *relevant information*, $\mathrm{I}[Z;Y]$ [SST10].

[SST10] Shamir et al., 'Learning and generalization with the information bottleneck'.

Definition 7.2. The *IB Curve* or *relevance-compression function* is the functional that expresses the IB problem [GBNT03]:

[GBNT03] Gilad-Bachrach et al., 'An Information Theoretic Tradeoff between Complexity and Accuracy'.

$$\mathrm{R}^{(\mathrm{IB})}(\mathrm{I}_Y) = \min_{p(z|x):\ \mathrm{I}[Z;Y] \geqslant \mathrm{I}_Y} \mathrm{I}[Z;X], \tag{7.7}$$

or alternatively:

$$\mathrm{I}_Y^{(\mathrm{IB})}(\mathrm{R}) = \max_{p(z|x):\ \mathrm{I}[Z;X] \leqslant \mathrm{R}} \mathrm{I}[Z;Y], \tag{7.8}$$

where the random variables form a Markov chain $Y \leftrightarrow X \leftrightarrow Z$ and the minimisation is over all the normalised conditional distributions $p(z|x) | \sum_x p(z|x) = 1$ for which the constraint is satisfied.

A straightforward observation is that the Markovian relation characterises $p(z)$ and $p(y|z)$ through [Slo02]

[Slo02] Slonim, 'The information bottleneck: Theory and applications'.

$$\begin{cases} p(z) = \sum_{x,y} p(x,y,z) = \sum_x p(x)p(z|x) \\ p(y|z) = \frac{1}{p(z)} \sum_x p(x,y,z) = \frac{1}{p(z)} \sum_x p(x,y)p(z|x). \end{cases} \tag{7.9}$$

7.3.1 *The information plane*

Moreover, the plane where the horizontal axis corresponds to $\mathrm{I}[Z;X]$ and the vertical axis to $\mathrm{I}[Z;Y]$, named **information plane** (see Figure 7.4) is the natural equivalent to the distortion-compression plane

in Rate-Distortion Theory (Figure 7.2). Let the pair R, I_Y denote some levels of compression and relevant information, respectively. If this pair is located below the curve, some compressed representation Z has a compression level $R = I[Z;X]$ and relevant information $I_Y = I[Z;Y]$. The points laying on the IB Curve are the optimal representations for a certain level of relevant-information (or precision) I_Y or a certain level of compression (or complexity) R.

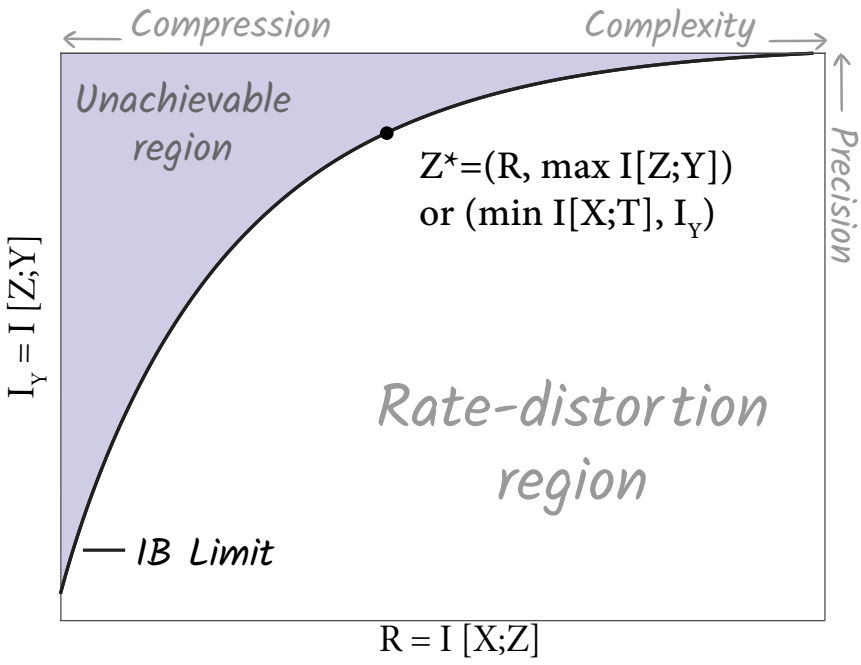


FIGURE 7.4: The IB Curve, $R^{(IB)}(D)$, in the information plane. Inspired by 'An Information Theoretic Tradeoff between Complexity and Accuracy' [GBNT03].

7.4 THE IB LAGRANGIAN

The Lagrangian relaxation of the IB functional is also a variational problem:

$$\mathcal{L}_\beta^{(IB)}[p(z|x)] = I[Z;X] - \beta I[Z;Y], \tag{7.10}$$

where β is the Lagrangian multiplier attached to the constrained relevant information [TPB99].

[TPB99] Tishby et al., 'The Information Bottleneck Method'.

At $\beta = 0$, no feature of the signal is relevant, and all messages are quantised (compressed) to a single point. At $\beta = \infty$, the solution is pushed toward arbitrarily detailed quantisation (no compression). 'By varying the (only) parameter, β, one can explore the tradeoff between the preserved meaningful information and compression at various resolutions' [TPB99].

Unlike the RDT problem (Section 7.1.3), in the IB problem, the constraint on the meaningful information is *nonlinear* in the mapping

$p(z|x)$, and it is a much harder variational problem. Notably, there is an analytical solution for IB Lagrangian (Eq. (7.10)). However, for the sake of clarity, before deriving this exact solution, we will show how IB can be seen as a particular case of RDT. This development will further help us to derive the analytical solution more directly.

7.5 IB PROBLEM AS A PARTICULAR CASE OF THE RDT PROBLEM

From the Data Processing Inequality (DPI) (Section 5.6.6),

$$\mathrm{I}[X;Y] \geqslant \mathrm{I}[Z;Y]. \tag{7.11}$$

Therefore, we can consider that the relevant information of X not captured by the representation Z is a natural choice for the expected distortion, as it represents a distortion in bits.

$$\langle d[x;z] \rangle = \mathrm{I}[X;Y] - \mathrm{I}[Z;Y] \geqslant 0 \tag{7.12}$$

From this definition, we can derive the following theorem:

Theorem 7.2. *If* $\langle d[x;z] \rangle_{p(x,z)} = \mathrm{I}[X;Y] - \mathrm{I}[Z;Y]$, *then* $d[x;z] = \mathrm{D}_{\mathrm{KL}}(p(y|x) \,||\, p(y|z))$.

Proof.

$$\begin{aligned} &\langle d[x;z] \rangle_{p(x,z)} = \mathrm{I}[X;Y] - \mathrm{I}[Z;Y] \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} - \sum_{z,y} p(z,y) \log \frac{p(z,y)}{p(z)p(y)}. \end{aligned} \tag{7.13}$$

Since $p(a,b) = p(b|a)p(a)$, we have:

$$= \sum_{x,y} p(y|x)p(x) \log \frac{p(y|x)\cancel{p(x)}}{\cancel{p(x)}p(y)} - \sum_{z,y} p(y|z)p(z) \log \frac{p(y|z)\cancel{p(z)}}{\cancel{p(z)}p(y)}. \tag{7.14}$$

From Eq. (7.9) :

$$= \sum_{x,y} p(y|x)p(x) \log \frac{p(y|x)}{p(y)} - \sum_{z,y,x} \frac{p(y|x)p(z|x)p(x)\cancel{p(z)}}{\cancel{p(z)}} \log \frac{p(y|z)}{p(y)} \tag{7.15}$$

$$= \sum_{x,y} p(y|x)p(x) \log \frac{p(y|x)}{p(y)} - \sum_{z,y,x} p(y|x)p(z,x) \log \frac{p(y|z)}{p(y)}. \tag{7.16}$$

From the normalisation constraint, $\sum_z p(z|x) = 1$:

$$= \cancel{1}^{\sum_z p(z|x)} \cdot \sum_{x,y} p(x)p(y|x) \log \frac{p(y|x)}{p(y)} - \sum_{z,y,x} p(y|x)p(z,x) \log \frac{p(y|z)}{p(y)} \tag{7.17}$$

$$= \sum_{z,x} p(z|x)p(x) \left[\sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \right] - \sum_{z,x} p(x,z) \left[\sum_y p(y|x) \log \frac{p(y|z)}{p(y)} \right] \tag{7.18}$$

$$= \sum_{z,x} p(x,z) \left[\sum_y p(y|x) \left(\log \frac{p(y|x)}{p(y)} - \log \frac{p(y|z)}{p(y)} \right) \right] \tag{7.19}$$

$$= \sum_{z,x} p(x,z) \left[\sum_y p(y|x) \left(\log \frac{p(y|x) \cancel{p(y)}}{\cancel{p(y)} p(y|z)} \right) \right] \tag{7.20}$$

$$= \mathbb{E}_{p(z,x)} \mathrm{D}_{\mathrm{KL}}(\, p(y|x) \, || \, p(y|z) \,). \tag{7.21}$$

Therefore

$$\langle d[x;z] \rangle_{p(x,z)} = \langle \mathrm{D}_{\mathrm{KL}}(\, p(y|x) \, || \, p(y|z) \,) \rangle_{p(x,z)} \tag{7.22}$$

$$d[x;z] = \mathrm{D}_{\mathrm{KL}}(\, p(y|x) \, || \, p(y|z) \,) \tag{7.23}$$

□

7.6 INFORMATION BOTTLENECK SOLUTION

Theorem 7.1 characterises the general form of the optimal solution to the rate-distortion problem. As we showed that the IB problem could be seen as a particular case of the RDT problem, the IB solution is straightforward:[12]

[12] The analytical solution to the IB problem is sometimes called the self-consistent equations.

Theorem 7.3. *The analytical solution of the variational problem*

$$\frac{\partial \mathcal{L}_\beta^{(IB)}[p(z|x)]}{\partial p(z|x)} = 0, \tag{7.24}$$

for normalised distributions $p(z|x)$ is given by the exponential form

$$\begin{cases} p(z|x) & = \frac{p(z)}{\mathrm{Z}(x,\beta)} \exp(-\beta \, \mathrm{D}_{\mathrm{KL}}(\, p(y|x) \, || \, p(y|z) \,)), \\ p(z) & = \sum_{x,y} p(x,y,z) = \sum_x p(x)p(z|x) \\ p(y|z) & = \frac{1}{p(z)} \sum_x p(x,y,z) = \frac{1}{p(z)} \sum_x p(x,y)p(z|x). \end{cases} \tag{7.25}$$

where Z is the normalisation factor (partition function). The Lagrange multiplier β is positive and

$$\beta = \frac{\partial \mathrm{I}[Z;Y]}{\partial \mathrm{I}[Z;X]}. \tag{7.26}$$

Proof. Apply $d[x;z] = \mathrm{D}_{\mathrm{KL}}(\, p(y|x) \, || \, p(y|z) \,)$ to Theorem 7.1. □

7.7 CONCLUDING REMARKS

In this section, we presented the Information Bottleneck (IB) Problem (Section 7.2.1) and the IB Lagrangian (Section 7.4), with its corresponding analytical companion, the information plane (Section 7.3.1).

The exciting aspect of the IB Problem is that it uses the "help" of a relevant variable to define the distortion measure. Therefore, we have a task-specific distortion measure (loss function). In opposition to MLT and RDT, which are loss-function-agnostic, in the IB method, the Kullback-Leibler divergence (D_{KL}) of the true distribution p and the model q emerges as the natural choice (Section 7.5).

Despite the similarities with the supervised learning problem (Section 4.2.1), the IB Problem assumes knowledge of the distribution $P(X, Y)$, and it is not yet in the realm of Information-Theoretical Machine Learning.

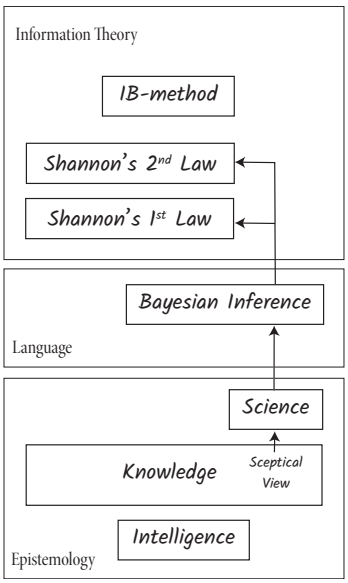


FIGURE 7.5: The IB method uses Shannon laws to define an unreacheable compression region in the information-plane.

8

Information Bottleneck and Representation Learning

'We know the past, but cannot control it.
We control the future, but cannot know it.'
—Claude Shannon

This chapter presents the idea of using the IB principle for representation learning in general, not specific for Deep Learning, which will be the subject of Chapter 9.

In Section 8.1, we show *why* to learn representations. In Section 8.2, we discuss *how* to characterise a good representation. Section 8.5 presents the two levels of representation in learning, which will help us understand *what* to represent.

In Section 8.3, we finally present the IB Learning problem, its difference to the IB method, *how* to find good representations with it, and its strengths and weaknesses as a representation learning framework. Finally, we close the chapter with Section 8.9 that brings evidence that the IB framework can predict bounds on human learning.

8.1 REPRESENTATION LEARNING

In our human experience, we know that a good representation of data is crucial for accomplishing tasks. The Hindu-Arabic numeral system advantages, for example, are so manifest that it has been adopted almost everywhere.

In the history of Machine Learning, good representations have always played a central role. In its first years, before trying to solve a task, researchers would *feature engineer*: use their knowledge of the problem in hand to *encode* the data into a representation easier for computers to learn the task.

The goal of designing features is to separate explanatory factors of variation behind the high dimensional observed data. The challenge is that many of the "factors of variation" influence every piece of data

[GBC16] Goodfellow et al., *Deep Learning*.

we can observe [GBC16].

Consider the problem of Object Classification[1]; each pixel depends on different factors: the viewing angle of the picture, the object's pose, the quality and calibration of the lens, the conditions of lightning, unrelated background objects.

[1] *"What object does this picture represent?"* Object Classification is the task of assigning a category (a label) for an image.

Over time, it became clear that the success of machine learning was so heavily dependent on appropriate features that finding them should also be part of the process of learning itself. Therefore, **representation learning** or **feature learning** is a set of techniques that allows a machine to learn features and use them to perform a specific task. Learned representations often result in better performance and flexibility, allowing a more straightforward adaptation of an AI system to new tasks, with the minimal human intervention [GBC16]. Furthermore, the recent success of Deep Learning, which is one of many ways to learn representations, has shown the power of this *encoder-decoder* scheme.

$$\underset{\text{input}}{X} \xrightarrow{\text{encoder}} \underset{\text{features}}{Z} \xrightarrow{\text{decoder}} \underset{\text{output}}{\hat{Y}}$$

FIGURE 8.1: Alessandro Achille.

[AS18a] Achille and Soatto, 'Emergence of Invariance and Disentangling in Deep Representations'.

8.2 DESIDERATA FOR REPRESENTATIONS

What are good representations of the data? A good representation makes a subsequent learning task easier [GBC16]. Achille and Soatto [AS18a] present a mathematical definition using information theory.

task : In supervised learning, we want to find the stochastic conditional distribution $p(y|x)$ of a target variable Y that we refer as the task:

$$Y := \mathrm{P}(Y|X = x)$$

representation : Z is a representation of X if it can be fully described by the stochastic conditional $p(z|x)$:

$$Z := \mathrm{P}(Z|X = x)$$

sufficient: Z is a sufficient representation of X w.r.t Y if $Y \to X \to Z$ form a Markov chain[2] and:

$$\mathrm{I}[Z;Y] = \mathrm{I}[X;Y].$$

minimal: Z has the smallest amount of information among all the sufficient representations of X. This means there is an encoding from X to Z that keeps only relevant information[3]:

$$\exists\, X \mapsto Z \mid \; \mathrm{I}[Z;X] = \mathrm{I}[Z;Y] = \mathrm{I}[X;Y] \tag{8.1}$$

[2] Note that $Y \neq \hat{Y}$. Here, the Markov chain is from the unknown target variable Y to the representation Z through the input. See Figure 8.4.

[3] Minimal representations are generally equated to low-dimensional data. However, as we have exposed in Chapter 5, a high dimensional representation can have little information. Thus, for example, sparse representations that force most of its bits to be zero are high-dimensional low-informational representations.

invariant: to the effect of nuisances (noise).[4] Let η be a nuisance for the task Y. If η does not have information about Y, there should not be information of η in the representation Z , the classifier could fit spurious correlations:

$$\begin{aligned} \eta \perp Y \Rightarrow \mathrm{I}[\eta; Y] &= 0 \\ \Rightarrow \mathrm{I}[Z; \eta] &= 0 \end{aligned} \tag{8.2}$$

[4] Nuisances are factors of variation that affect data, but are otherwise irrelevant for the task.

maximally disentangled: information lies on components of the representation Z and not in the correlations of them. Then, mathematically, let TC denote the total correlation, *a.k.a. multi-information*.

$$\mathrm{TC}(Z) = \mathrm{D}_{\mathrm{KL}}(p(z) || \prod_{i}^{n} p(z_i)), \tag{8.3}$$

$$\mathrm{TC}(Z) = 0 \implies z_1 \perp z_2 \perp \cdots \perp z_n. \tag{8.4}$$

This desiderata for representations corresponds directly with our goals for learning algorithms. We want our models to predict the task correctly (sufficiency). Simultaneously, we want them to generalise to out-of-sample examples (invariance to nuisance factors).

$$\begin{aligned} \text{accuracy} &\leftrightarrow \text{sufficiency} \\ \text{generalisation} &\leftrightarrow \text{invariance/minimality} \\ \text{explainability} &\leftrightarrow \text{disentanglement} \end{aligned}$$

FIGURE 8.2: Correspondence of desired properties of learning algorithms and representations.

Another desired characteristic, albeit often forgotten, is that we want our models to be explainable.[5] This characteristic relates to *disentangling* the underlying causes (factors) of the observed data (maximally disentangled) [GBC16].

[5] Disentanglement and minimality also simplify the subsequent inference (decoding).

Although disentanglement may be an abstract characteristic not very well defined, Achille and Soatto [AS18a] propose a simplification by defining it as the total correlation of the representation features.

The only property of the desiderata that still does not correspond with learning algorithms is *minimality*. However, it is straightforward that a small sufficient representation has a smaller chance of containing spurious correlations, and it is more likely to generalise well. Minimal sufficient representations have no spurious factors that do not explain the variability of the observed data. As we will show, a representation is invariant only if it is also minimal.

8.2.1 *Invariant if minimal*

Theorem 8.1 ([Ach19], Proposition 2.4.1)**.** *Let* η *be a nuisance for the target* Y *and let* Z *be a sufficient representation of the input* X *w.r.t* Y*. Suppose that* Z *depends on* η *only through* X *(i.e. ,* $\eta \to X \to Z$*). We also consider that* X *has all information about* Y*; therefore, we can say that it is a* ***deterministic function*** *of* Y *and nuisances* $X := f(Y;\eta)$*.*

To say that Z *is invariant if and only if it is minimal implies that* $\mathrm{I}[Z;\eta] = \mathrm{I}[Z;X] - \mathrm{I}[X;Y]$*:*

$$\begin{aligned} \forall\, Z \mid \mathrm{I}[Z;Y] &= \mathrm{I}[X;Y], \\ \mathrm{I}[Z;X] = \mathrm{I}[X;Y] &\iff \mathrm{I}[Z;\eta] = 0, \\ \mathrm{I}[Z;\eta] &= \mathrm{I}[Z;X] - \mathrm{I}[X;Y]. \end{aligned}$$

This equality holds up to a small residual ϵ*:*

$$\mathrm{I}[Z;\eta] = \mathrm{I}[Z;X] - \mathrm{I}[X;Y] - \epsilon,\ 0 \leqslant \epsilon \leqslant \mathrm{H}[Y|X] \tag{8.5}$$

Proof.

$$\begin{cases} Y, \eta \to X \to Z & \text{(by definition)} \\ \mathrm{I}[Z;Y,\eta] \leqslant \mathrm{I}[Z;X] & \text{(DPI)} \\ \mathrm{I}[Z;Y,\eta] = \mathrm{I}[Z;\eta] + \mathrm{I}[Z;Y|\eta] & \text{(chain rule)} \end{cases}$$

$$\begin{aligned} &\mathrm{I}[Z;\eta] + \cancelto{\mathrm{I}[Z;Y]}{\mathrm{I}[Z;Y|\eta]} \leqslant \mathrm{I}[Z;X] && (\eta \perp Y) \\ &\mathrm{I}[Z;\eta] \leqslant \mathrm{I}[Z;X] - \cancelto{\mathrm{I}[X;Y]}{\mathrm{I}[Z;Y]} && (Z \text{ sufficiency}) \\ &\mathrm{I}[Z;\eta] = \mathrm{I}[Z;X] - \mathrm{I}[X;Y] - \epsilon, \epsilon \geqslant 0 && (\epsilon \text{ lower bound}) \end{aligned}$$

Now we only need to prove the upper bound for ϵ:

$$\begin{aligned} \epsilon &= \mathrm{I}[Z;X] - \mathrm{I}[Z;\eta] - \mathrm{I}[X;Y] \\ &= \mathrm{I}[Z;Y,\eta] - \mathrm{I}[Z;\eta] - \mathrm{I}[X;Y] && (X := f(Y;\eta)) \\ &= \cancel{\mathrm{I}[Z;\eta]} + \mathrm{I}[Z;Y|\eta] - \cancel{\mathrm{I}[Z;\eta]} - \mathrm{I}[X;Y] && \text{(chain rule)} \\ &= \cancelto{\mathrm{H}[Y]}{\mathrm{H}[Y|\eta]} - \mathrm{H}[Y|\eta;Z] - \mathrm{H}[Y] + \mathrm{H}[Y|X] && (\eta \perp Y) \\ &= \cancel{\mathrm{H}[Y]} - \mathrm{H}[Y|\eta;Z] - \cancel{\mathrm{H}[Y]} + \mathrm{H}[Y|X] \\ &\leqslant \mathrm{H}[Y|X] && (\epsilon \text{ upper bound}) \end{aligned}$$

$$\mathrm{I}[Z;\eta] = \mathrm{I}[Z;X] - \mathrm{I}[X;Y] - \epsilon,\ 0 \leqslant \epsilon \leqslant \mathrm{H}[Y|X] \qquad \square$$

As a consequence of this proposition, it is possible to construct invariant representations, which will generalise well, by reducing the

amount of information the representation Z contains about the input X while keeping $\mathrm{I}[Z;Y]$, the amount of information we need for the task. As $\mathrm{H}[Y] \ll \mathrm{H}[X]$ **the compressibility of the input determines generalisation**. Thus, it is independent on the hypothesis space of the learning algorithm.

8.3 IBT LEARNING PROBLEM: LEARNING APPROXIMATELY MINIMAL SUFFICIENT DISENTANGLED REPRESENTATIONS

We have discussed what constitutes a good representation. This section is about finding such representations. For that, we will adjust the IB Problem Setting (Section 7.2.1) for supervised learning.[6]

[6] For consistency with Chapter 4, we will repeat some definitions in this section.

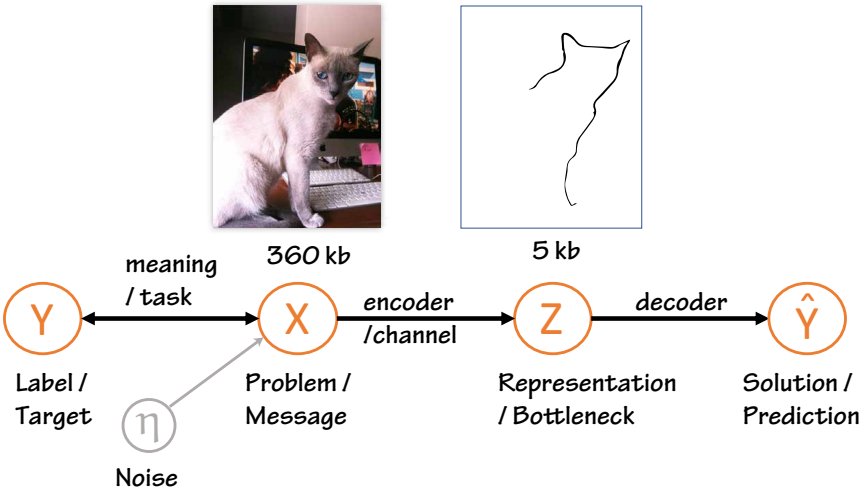


FIGURE 8.3: The IBT Learning Problem is the adaptation of the IB Problem to the learning setting.

8.3.1 *Definitions*

1. Let X be the random variable that denotes the **generator** (or **source**) of instance vectors x of the learning problem (**messages**), randomly drawn from a probability distribution $\mathrm{P}(X)$, $x \sim \mathrm{P}(X),\ x \in \mathbb{A}_X$;

2. Let Y be a random **relevant variable** (the **Target**) which represents the solution y for the problem x, *i.e.* the intended meaning $p(y|x)$ of the message x, $y \sim \mathrm{P}(Y),\ y \in \mathbb{A}_Y$;

3. A **task supervisor** knows the **task** distribution $\mathrm{P}(Y|X)$ and returns an output vector y_i for every input vector $\boldsymbol{x}_i$[7]: $\boldsymbol{y}_i := p(y|\boldsymbol{x}_i)$;

[7] Notice that here $\boldsymbol{y}_i$ is not the label but a vector that represents the probability of each label.

4. Let Z be a **bottleneck** random variable that denotes a compressed representation of the input X that is sufficient w.r.t. Y and obeys the Markov chain $Y \leftrightarrow X \leftrightarrow Z$;

5. Let the stochastic conditional distribution $q(z|x)$ be an **encoder**of input instances into representations,
$z := q(z|x)$.

6. Let the stochastic conditional distribution $q(y|z)$ be a **decoder** of representations into solutions of the problem,
$\hat{y} := q(y|z)$.

7. A **learning algorithm** $\mathcal{A}$, which is the functional that given a dataset $\mathrm{D}_n = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ of n inputs and outputs of the task, selects a hypothesis $h = \underset{\text{decoder}}{q(y|z)} \circ \underset{\text{encoder}}{q(z|x)}$ from the hypothesis space $\mathcal{H}$:

$$\mathcal{A} : \underbrace{(\mathcal{X} \times \mathcal{Y})^n}_{\mathrm{D}_n} \to \mathcal{H}. \tag{8.6}$$

8.3.2 *Assumptions*

i. The random variables X, Y and Z, are **discrete**;

ii. $Y \to X \to Z$ form a Markov-chain;

iii. $\mathbb{A}_X$, $\mathbb{A}_Y$ and $\mathbb{A}_Z$ are **finite sets**;

iv. **No assumption on** $\mathrm{D} = \mathrm{P}(X, Y)$.

v. $\mathrm{D} = \mathrm{P}(X, Y)$ **is unknown at the training stage**.

vi. $\mathrm{D} = \mathrm{P}(X, Y)$ **is fixed:** the ordering of examples in the sample is irrelevant.

vii. X is i.i.d. sampled.[8]

viii. The encoder and the decoder are **stochastic** mappings.[9]

ix. the **distortion measure** between X and its representation Z is $\mathrm{I}[X;Y] - \mathrm{I}[Z;Y] = \mathrm{D}_{\mathrm{KL}}(p(y|x)||p(y|z))$.[10]

x. the **entanglement** of a random variable Z is defined as *total correlation* of its components [AS18b].

[8] We could use an ergodic process, but for simplification we will use i.i.d. sampling.

[9] Notice that given the Markov chain $Y \leftrightarrow X \leftrightarrow Z$, due to reparemetrisation invariance (Theorem 5.4 (reparametrisation invariance (RI))), a deterministic mapping of the data does not throw out information, i.e. let $f : \mathbb{A}_X \to \mathbb{A}_Y$ be deterministic, $\mathrm{I}[f(X); Y] = \mathrm{I}[X; Y]$.

[10] This assumption is not strictly required, as it can be derived. The only reason to keep it here is to make the comparison of different problem settings easier.

[AS18b] Achille and Soatto, 'Information Dropout: Learning Optimal Representations Through Noisy Computation'.

8.3.3 *Problem statement*

Given the problem setting above, the IBT learning problem is to find the encoder $p(z|x)$ and decoder $p(y|z)$ such that:

1. the encoder maximises the compression of the input X into the representation Z while preserving the maximum information about the "meaning" Y. In other words, the encoder that generates minimal sufficient disentangled representations of the input.
2. the decoder is trivial as a result of the characteristics of the representation.
3. The selection is based on a training set of n i.i.d. observations drawn from the distribution $\mathrm{P}(X, Y)$.

8.3.4 *IBT learning as a variational problem*

Finding the encoder for minimal sufficient disentangled representations is equivalent to finding a distribution $p(z|x)$ that solves the following constrained optimisation problem:

$$\begin{aligned} q(z|x) := \arg\min_{p(z|x)} \quad & \mathrm{I}[Z;X] \\ \text{s.t.} \quad & 0 \leqslant \mathrm{I}[X;Y] - \mathrm{I}[Z;Y] \\ & 0 \leqslant \mathrm{TC}(Z). \end{aligned} \tag{8.7}$$

This nonlinearly constrained optimisation problem[11] is very similar to the IB Problem (Section 7.2). It just adds the total correlation constraint and assumes no knowledge over $\mathrm{P}(X, Y)$. Tishby et al. [TPB99] proposed solving the IB problem using a relaxed minimisation, the IB Lagrangian:

$$\begin{aligned} \min_{p(z|x)} \quad & \mathrm{I}[Z;X] \\ \text{s.t.} \quad & \mathrm{I}[Z;Y] \leqslant \mathrm{I}[X;Y] \end{aligned} \quad \Longrightarrow \quad \begin{aligned} & \min \mathrm{I}[Z;X] + \beta(\cancel{\mathrm{I}[X;Y]} - \mathrm{I}[Z;Y]), \\ & \min \mathrm{I}[Z;X] - \beta \mathrm{I}[Z;Y]. \end{aligned} \tag{8.8}$$

Let us also apply a Lagrangian relaxation to our representation

[11] Prior to the publishing of [Ale+16], there was no known algorithm to minimise the IB Lagrangian for discrete X and Y with large state spaces or non-Gaussian continuous joint distribution.

[TPB99] Tishby et al., 'The Information Bottleneck Method'.

learning problem:

$$\mathcal{L} = \mathrm{I}[Z;X] + \beta(\mathrm{I}[X;Y] - \mathrm{I}[Z;Y]) + \gamma \mathrm{TC}(z), \tag{8.9}$$

$$\text{Let } \beta^{-1} = \frac{1}{\beta}, \tag{8.10}$$

$$\gamma' = \frac{\gamma}{\beta} \tag{8.11}$$

$$\mathcal{L} = (\mathrm{I}[X;Y] - \mathrm{I}[Z;Y]) \overset{\mathrm{H}[Y|Z]}{\rightarrow} + \beta^{-1}\mathrm{I}[Z;X] + \gamma' \mathrm{TC}(z), \tag{8.12}$$

$$\mathcal{L} = \mathrm{H}[Y|Z] + \beta^{-1}\mathrm{I}[Z;X] + \gamma' \mathrm{TC}(z). \tag{8.13}$$

Let us denote $q_\theta(z|x)$ (encoder) and $q_\theta(y|z)$ (decoder) the unknown conditional distributions we want to estimate[12], parametrised by θ.

[12] $q_\theta(y|x) = \sum q_\theta(z|x) q_\theta(y|z)$
$q(y|x, \theta^*) = p(y|x)$

Then, rewriting the Lagrangian as a per sample loss function, we have:

$$\mathrm{H}[Y|Z] \approx \mathbb{E}_{(x,y)\sim p(x,y)}[\mathbb{E}_{z\sim q_\theta(z|x)} - \log q_\theta(y|z)] \tag{8.14}$$

$$\mathrm{I}[Z;X] = \mathbb{E}_{x\sim p(x)} \mathrm{D}_{\mathrm{KL}}(q_\theta(z|x) || p(z)) \tag{8.15}$$

$$\mathrm{TC}(z) = \mathrm{D}_{\mathrm{KL}}(p(z) || \textstyle\prod_j q(z_j)) \tag{8.16}$$

$$\begin{aligned}\hat{\mathcal{L}} = \frac{1}{n}\sum^{n} & \mathbb{E}_{z\sim q_\theta(z|x_i)} - \log q_\theta(y_i|z) \\ & + \beta^{-1}\mathrm{D}_{\mathrm{KL}}(q_\theta(z|x_i) || p(z)) \\ & + \gamma' \mathrm{D}_{\mathrm{KL}}(p(z) || \textstyle\prod_j q_\theta(z_j)).\end{aligned} \tag{8.17}$$

The second and third terms of the loss are intractable, as we need to know $p(z)$ to compute, which is an unknown of our problem. Achille and Soatto, however, prove that if $\beta^{-1} = \gamma'$, *i.e.* we assume a factorised unknown distribution, the Lagrangian can be solved [AS18b].

[AS18b] Achille and Soatto, 'Information Dropout: Learning Optimal Representations Through Noisy Computation'.

$$\hat{\mathcal{L}} = \underbrace{\frac{1}{n}\sum^{n} \mathbb{E}_{q_\theta(z|x_i)} - \log q_\theta(y_i|z)}_{\hat{\mathrm{H}}(p, p_\theta)} + \beta^{-1}\mathrm{D}_{\mathrm{KL}}(q_\theta(z|x_i) || q_\theta(z)), \tag{8.18}$$

$$q_\theta(z) = \prod_j q_\theta(z_j). \tag{8.19}$$

$$\hat{\mathcal{L}} = \hat{\mathrm{H}}(p, p_\theta) + \beta^{-1}\mathrm{D}_{\mathrm{KL}}(q_\theta(z|x) || q_\theta(z)) \qquad \text{Activations IB} \tag{8.20}$$

Where $\hat{\mathrm{H}}(p, p_\theta)$ is the cross-entropy, and the second term is a regulariser that penalises the transfer of information from X to Z. In other words, the regulariser penalises complexity measured as $\mathrm{I}[Z;X]$. The usage of cross-entropy loss and this kind of regularisers is

widespread in practice. Nevertheless, Achille gave theoretical ground for such choices[13] [Ach19].

Minimising the standard IB Lagrangian assuming the activations are independent, i.e. $q(z) = \prod_i q(z_i)$ is equivalent to enforcing disentanglement. Practitioners already adopt this independence assumption on the grounds of simplicity since the actual marginal $p(z)$ is incomputable. Higgins et al. also empirically observed that using a factorised model results in "disentanglement" [Hig+17]. Because of the previous propositions, we can assume the activations are indeed independent and ignore the TC term.[14]

Corollary 1. *Any learning algorithm that:*

- *assumes a stochastic $p(y|x)$;*
- *uses a D_{KL}-equivalent loss (for example the cross-entropy loss or the logistic loss);*
- *and a regularisation term that penalises the amount of information of the input stored in the model,*

is learning a minimal sufficient disentangled representation and, in fact, solving the IB learning problem.

8.4 RETHINKING GENERALISATION: CROSS-ENTROPY AND OVERFITTING

In previous sections, we derived the cross-entropy loss (Eq. (8.20)) from a list of desired properties for representations(Section 8.2). We also showed that generalisation relates to the compressibility of the input(Section 8.2.1).

Zhang et al. demonstrates that the expressivity of DNNs is enough to fit random labels [Zha+16]. Thus, at least for DNNs, generalisation is more *not overfitting* than not underfitting. This characteristic may be the case for other learning techniques as well. In this section, we will keep *rethinking* generalisation on this new information-theoretical perspective and try to elucidate how cross-entropy loss relates to overfitting and memorisation.

Classical MLT assumes that we select a hypothesis h parametrised by θ. Conceptually, we already *rethought* generalisation as determined only by the compressibility of the input (Section 8.2.1). In this sense, the task is determined by the training dataset only.[15] Thus, instead of a parametrised model, we will assume a parametrised unknown distribution $P(D|\theta)$. In this context [AS18a]:

[13] The reference constraint their findings to DNNs optimised with SGD. We regard the result more general than that.

[Ach19] Achille, 'Emergent Properties of Deep Neural Networks'.
URL: https://escholarship.org/uc/item/8gb8x6w9

[Hig+17] Higgins et al., 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework'.

[14] This insight allowed Alemi et al.; Achille and Soatto independently develop basically the same algorithm for estimating mutual information for any distribution using DNNs [Ale+16; AS18b].

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization.*

[15] Achille defines the task as the dataset distribution for which we only have one sample [Ach19].

[AS18a] Achille and Soatto, 'Emergence of Invariance and Disentangling in Deep Representations'.

Theorem 8.2. *Given* $D = (X, Y)$, $D \sim \mathrm{P}(X, Y|\theta)$, *and a representation* W *of* θ, *s.t.* $Y|X \leftrightarrow \mathrm{W} \leftrightarrow \theta$ *form a Markov-chain:*

$$\mathrm{H}_{p,q}[D|W] = \mathrm{H}_p[D, \theta] + \mathrm{I}[\theta; D|W] + \mathrm{D}_{\mathrm{KL}}(p \parallel q) - \mathrm{I}[D; W|\theta]$$

Proof. Notice that the output weight W of the training process can be seen as a random variable (that depends on the stochasticity of the initialisation, training steps, and the data); i.e. W is a representation of the dataset D and we can talk about $\mathrm{I}[W; D]$.

First, we show that minimising $\mathrm{H}_{p,q}[y|x]$ is equivalent to minimising $\mathrm{H}_{p,q}[x, y]$

$$min\mathrm{H}_{p,q}[y|x, w] = min\mathrm{H}_{p,q}[D|W]. \tag{8.21}$$

When a learning algorithm optimises the cross-entropy loss, it is effectively just minimising the KL-divergence, as the first term (entropy) is a constant:

$$min\mathrm{H}_{p,q}[y|x, w] = min\left(\cancel{\mathrm{H}_p[y|x, w]} + \mathbb{E}\mathrm{D}_{\mathrm{KL}}(p(y|x, \theta)||q(y|x, w))\right). \quad \text{from (5.} \tag{8.22}$$

The same happens in the minimisation of the cross-entropy of the joint dataset:

$$min\mathrm{H}_{p,q}[x, y|w] = min\left(\cancel{\mathrm{H}_p[x, y, w]} + \mathrm{D}_{\mathrm{KL}}(p(x, y, \theta)||q(x, y, w))\right) \quad \text{from (5.3} \tag{8.23}$$

Here, we show that the divergence of $y|x$ is the same as the divergence of joint distribution x, y, a step that was assumed by Achille and Soatto:

$$\mathrm{D}_{\mathrm{KL}}(p(x, y)||q(x, y)) = \mathbb{E}_p \log \frac{p(x, y)}{q(x, y)} \tag{8.24}$$

$$= \mathbb{E}_p \log \frac{p(y|x)p(x)}{q(y|x)p(x)} = \mathbb{E}_p \left[\log p(y|x)p(x) - \log q(y|x)p(x)\right] \tag{8.25}$$

$$= \mathbb{E}_p \left[\log p(y|x) + \cancel{\log p(x)} - (\log q(y|x) + \cancel{\log p(x)})\right] \tag{8.26}$$

$$= \mathbb{E}_p \left[\log p(y|x) - \log q(y|x)\right] = \mathbb{E}_p \log \frac{p(y|x)}{q(y|x)} \tag{8.27}$$

$$= \mathrm{D}_{\mathrm{KL}}(p(y|x)||q(y|x)) \tag{8.28}$$

Therefore we can say that a learning algorithm minimises $\mathrm{H}_{p,q}[D|W]$.

$$\mathrm{H}_{p,q}[D|W] = \mathrm{H}_p[\mathrm{D}, \mathrm{W}] + \mathrm{D}_{\mathrm{KL}}(\mathrm{P}(D, \theta) \parallel \mathrm{Q}(D, W)) \quad \text{from (5.39)} \tag{8.29}$$

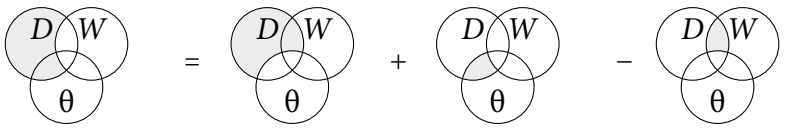
To prove that:

$$\mathrm{H}_{p,q}[D|W] = \mathrm{H}_p[D|\theta] + \mathrm{I}[\theta; D|W] + \mathbb{E}\mathrm{D}_{\mathrm{KL}}(p \parallel q) - \mathrm{I}[D; W|\theta],$$

we just need to prove that:

$$\mathrm{H}_p[D|W] = \mathrm{H}_p[D, \theta] + \mathrm{I}[D|W; \theta] - \mathrm{I}[D; W|\theta]. \tag{8.30}$$

This equivalence is clear with the help of the following Venn diagrams[16]:



$$\mathrm{H}_p[D|W] = \mathrm{H}_p[D|\theta] + \mathrm{I}[D|W; \theta] - \mathrm{I}[D; W|\theta]$$

□

Let us examine the cross-entropy decomposition:

$$\mathrm{H}_{p,q}[D|W] = \underbrace{\mathrm{H}_p[D|\theta]}_{\text{intrinsic error}} + \underbrace{\mathrm{I}[\theta; D|W]}_{\text{sufficiency}} + \underbrace{\mathrm{D}_{\mathrm{KL}}(p \parallel q)}_{\text{efficiency}} - \underbrace{\mathrm{I}[D; W|\theta]}_{\text{memorisation}}$$

intrinsic error: $\mathrm{H}_p[D|\theta]$ relates to the intrinsic error that we would find even if we knew p_θ;

sufficiency: $\mathrm{I}[\theta; D|W]$ measures how much information of θ was compressed in the weights;

efficiency: $\mathrm{D}_{\mathrm{KL}}(p \parallel q)$ measures the efficiency[17] of the representation, *i.e.* the number of additional bits we need to represent the input with $q(w|\mathrm{D})$ instead of using p_θ (see Section 5.5.4);

memorisation: $\mathrm{I}[D; W|\theta]$ is the last and only negative term. It relates to overfitting and measures how much information about the dataset unrelated to θ is memorised in the weights.

The optimiser will try to increase memorisation because it is the only negative term. Thus, Achille and Soatto propose a naïve method to eliminate this proneness to overfitting: adding back the memorisation term in the lossThus, [AS18a].

$$\mathcal{L}(W) = \mathrm{H}_{p,q}[D|W] + \mathrm{I}[D; W|\theta] \tag{8.31}$$

To calculate $\mathrm{I}[D; W|\theta]$ true distribution, p_θ is needed. Nevertheless, we are just trying to approximate p_θ with q during training. Hence

[16] Our assumptions guarantee that all information measures in the diagram are positive, thus there is no problem in using the Venn diagram in this case.

[17] It relates to generalisation, as additional bits of information can correlate to noise.

[AS18a] Achille and Soatto, 'Emergence of Invariance and Disentangling in Deep Representations'.

we are presented with the *chicken-egg* problem. Rather, one can add a Lagrangian multiplier to upper bound $\mathrm{I}[D; W|\theta]$:

$$\mathcal{L}(W) = \mathrm{H}_{p,q}[D|W] + \beta^{-1}\mathrm{I}[D; W] \qquad \text{Weights IB} \qquad (8.32)$$

Remarkably, this has the same form as the IB Lagrangian, Eq. (7.10). When $\beta = 1$, (8.32) reduces to the Evidence Lower Bound (ELBO) loss used in variational inference [Ach19, p. 53].

[Ach19] Achille, 'Emergent Properties of Deep Neural Networks'.
URL: https://escholarship.org/uc/item/8gb8x6w9

8.5 TWO LEVELS OF REPRESENTATION

Some criticism on IBT derive from a lack of rigour in explaining the fundamentals (see Section 9.3.4).

The crucial problem in MLT is that we want to predict the behaviour (bound) of learning algorithms in future data while we can only access past performance.

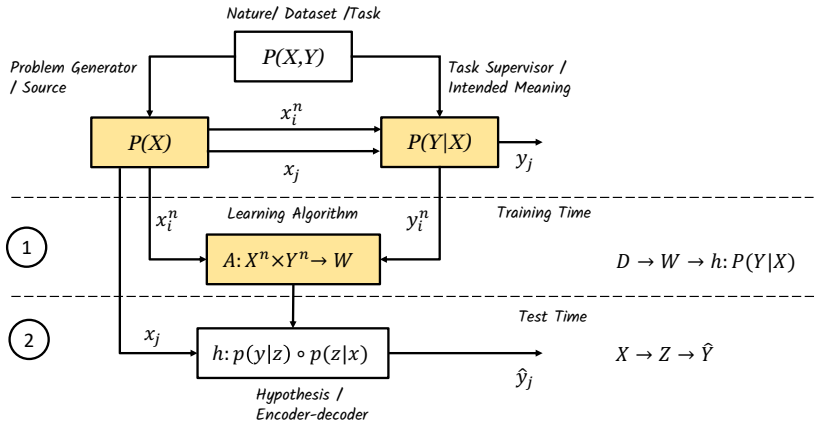


FIGURE 8.4: Two levels of representation in the learning setting.

This dichotomy translates to representation learning by two intertwined but different representations:

1. the representation of a *dataset* (past data), a function that can be stored in memory for the later accomplishment of the task. It needs to keep useful information for future decisions without squandering resources in remembering spurious correlations or one-time events.

2. the representation of an *input example* (current data): which need to keep the essence of the scene at hand;

Borrowing the terminology of Deep Learning, Achille; Achille and Soatto call these two levels of representation of *information in the weights* and *information in the activations*, respectively [Ach19; AS19].

[AS19] Achille and Soatto, *Where is the Information in a Deep Neural Network?*.

Several IBT papers do not address this difference. In particular, some of the seminal work [TZ15a; ST17; Tis17b]. How can we minimise the information in the activations while we cannot access future data? There is a missing step.

Notice that we now have two Lagrangians. The original (Section 8.3.4) and this new Lagrangian emerged from eliminating overfitting.

$$\underset{\text{input}}{X} \longrightarrow \underset{\text{activations}}{Z} \longrightarrow \underset{\text{label}}{Y}$$

$$\min_{q(Z|X)} \mathcal{L}(\mathrm{W}) = \mathrm{H}_{p,q}[Y|Z] + \beta^{-1}\mathrm{I}[Z;X] \qquad \text{Activations IB}$$

$$\underset{\text{dataset}}{\mathrm{D}} \longrightarrow \underset{\text{weights}}{\mathrm{W}} \longrightarrow \underset{\text{real distribution}}{\mathrm{P}(Y|X)}$$

$$\min_{q(D|W)} \mathcal{L}(\mathrm{W}) = \mathrm{H}_{p,q}[D|W] + \beta^{-1}\mathrm{I}[W;D] \qquad \text{Weights IB}$$

Intuitively, there is a strong connection between *information in the weights* and *information in the activations*. $\mathrm{I}[Z;X]$, 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 the optimal activation representation z^*, which minimises $\mathrm{I}[Z;X]$.

In 'Emergence of Invariance and Disentangling in Deep Representations', Corollary C.8, Achille and Soatto have proved that indeed there is a bound:

$$\mathrm{I}[Z;X] \leqslant \mathrm{I}[W;D] \tag{8.33}$$

As $\tilde{\mathrm{I}}[w;D]$ can be calculated, this development allows one to regularise the training explicitly. This explicit regularisation is what 'Information Dropout: Learning Optimal Representations Through Noisy Computation', Information Dropout proposes [AS18b].

Besides, even without calculating the information in the weights one can control it by injecting noise, which can be modulated from zero, no effect in the rate of the encoder, to the capacity of the channel, which leaves the encoder with no information left.

> *We know the past but cannot control it.*
> *We control the future but cannot know it.*
>
> — Claude Shannon

[TZ15a] Tishby and Zaslavsky, 'Deep learning and the information bottleneck principle'.

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

[Tis17b] Tishby, *Information Theory of Deep Learning*.
URL: https://youtu.be/FSfN2K3tnJU

[AS18b] Achille and Soatto, 'Information Dropout: Learning Optimal Representations Through Noisy Computation'.

8.6 SHANNON VS. FISHER INFORMATION

We still have the problem that to calculate $\mathrm{I}[X;Y]$, we need to know $\mathrm{P}(X,Y)$. We can, however, bound the amount of information using Fisher Information Section 5.8.2. We use:

$$\mathrm{I}[X;Y] = \mathrm{D}_{\mathrm{KL}}(\mathrm{P}(X,Y)\|\mathrm{P}(Y)\mathrm{P}(X)) \tag{8.34}$$
$$= \mathbb{E}_X \mathrm{D}_{\mathrm{KL}}(\mathrm{P}(Y|X)\|\mathrm{P}(Y)), \tag{8.35}$$

to rewrite Eq. (8.32)(Weights IB) as:

$$\mathcal{L}(\mathrm{W}) = \mathrm{H}_{p,q}(\mathrm{D}|\mathrm{W}) + \beta^{-1}\mathrm{D}_{\mathrm{KL}}\left(\underbrace{\mathrm{Q}(\mathrm{W}|\mathrm{D})}_{\text{training output}} \| \underbrace{\mathrm{P}(\mathrm{W})}_{\text{fixed prior}}\right)$$

In other words, $\mathrm{I}(\mathrm{W};\mathrm{D})$ is the divergence of the encoder $\mathrm{Q}(\mathrm{W}|\mathrm{D})$ and the expected prior averaging all possible datasets, *i.e.* the unknown distribution. If we change the assumption Section 8.3.2, Item iv, and assume that the unknown distribution is an isotropic Gaussian[18], the information in the weights when W_* is minimal, is given by:

$$\mathrm{D}_{\mathrm{KL}} q(\mathrm{W}|\mathrm{D}) p(\mathrm{W}) = \frac{1}{2}\left(\log|\mathrm{F}n| + \cancel{\log\lambda^2\mathrm{I}} + \frac{\mathrm{W}_*^2}{\cancel{\lambda^2\mathrm{I}}}\right),$$

where the cancelled terms are the ones that do not depend on $\mathrm{Q}(\mathrm{W}|\mathrm{D})$ and can be ignored, $\log|\mathrm{F}|$ is the log-determinant of Fisher Information Matrix of the weights, and n is the number of samples in the dataset.

This assumption is quite interesting as it gives us an analytical and fast calculation of a bound to $\mathrm{I}(\mathrm{W};\mathrm{D})$:

$$\mathrm{I}[Z;X] \leqslant \mathrm{I}[W;D] \leqslant \log|\mathrm{F}(\mathrm{W}^*)| \tag{8.36}$$

Even if the unknown distribution $\mathrm{P}(\mathrm{D})$ is not an isotropic gaussian, we can think that near optima, it approximates one. We can arrive at the same result by approximating the Hessian with a Taylor expansion.

[18] An isotropic Gaussian is one where the covariance matrix is represented by $\Sigma = \lambda^2\mathrm{I}$.

8.7 CONNECTION TO VARIATIONAL AUTOENCODERS

Achille and Soatto show how the previous development relate with Variational Auto-encoders (VAEs) [AS18b]. Variational Auto-encoders (VAEs) [KW14] aim to reconstruct, given a training dataset $\mathcal{D} = x_i$, a latent variable z. The paper proposes that this can be thought as generating z through some unknown generative process $p_\theta(x|z)$. In practice, this is done by minimising:

$$\mathcal{L} = \frac{1}{\mathrm{N}}\sum^{\mathrm{N}} \mathbb{E}_{z\sim p_\theta(z|x_i)} - \log p_\theta(y_i|z) + \mathrm{D}_{\mathrm{KL}}(p_\theta(z|x_i)\|\prod_i q_\theta(z_i)).$$

[AS18b] Achille and Soatto, 'Information Dropout: Learning Optimal Representations Through Noisy Computation'.

[KW14] Kingma and Welling, 'Auto-Encoding Variational Bayes'.
URL: http://arxiv.org/abs/1312.6114

This minimisation is performed through sampling using SVGB [KW14]. It is clear by the formulation that VAE is equal to Eq. (8.32)(Weights IB) where $\beta = 1$.

8.8 CONNECTION TO PAC-BAYES

Achille and Soatto also relate their work with PAC-Bayes [AS18b]. From [McA13, Thrm 2],

[McA13] McAllester, 'A PAC-Bayesian Tutorial with A Dropout Bound'.

$$\begin{aligned}&\forall(\text{fixed})\lambda > 1/2,\ p(w),\ q(w|\mathrm{D}),\\&\mathbb{E}_{\mathrm{D}}[\mathrm{L}^{\text{test}} q(w|\mathrm{D})] \leqslant\\&\frac{1}{\mathrm{N}(1-\frac{1}{2\lambda})}\left(\mathrm{H}_{p,q}(y|x,w) + \lambda \mathrm{L}_{\max}\mathbb{E}_{\mathrm{D}}[\mathrm{D}_{\mathrm{KL}}[q(w|\mathrm{D})||p(w)]]\right)\end{aligned} \tag{8.37}$$

where $\mathrm{L}_{\max}$ is the maximum per-sample loss function. The right-hand side (RHS) coincides, modulo a constant, with Eq. (8.32) if we use $q(w)$ instead of $p(w)$. Since

$$\begin{aligned}&\mathrm{E}_{\mathrm{D}}[\mathrm{D}_{\mathrm{KL}}[q(w|\mathrm{D})||q(w)]]\\=&\mathrm{E}_{\mathrm{D}}[\mathrm{D}_{\mathrm{KL}}[q(w|\mathrm{D})||p(w)]] - \mathrm{D}_{\mathrm{KL}}[q(w)||p(w)] && (8.38)\\\leqslant&\ \mathbb{E}_{\mathrm{D}}[\mathrm{D}_{\mathrm{KL}}(q(w|\mathrm{D})||p(w))], && (\text{PAC-Bayes})\end{aligned}$$

the sharpest PAC-Bayes upper bound to the test error is obtained when $p(w) = q(w)$, in which case, Eq. (PAC-Bayes) reduces (modulo a constant) to the IB Lagrangian of the weights. Unfortunately, the marginal $q(w)$ of the weights is not tractable, as already stated. To circumvent this problem, we consider instead that the sharpest PAC-Bayes upper bound that can be obtained using a tractable factorised prior $p(w) = \tilde{q}(w) = \prod_i q(w_i)$.[19]

[19] This assumption is also made in the MDL framework.

8.8.1 *Relation to Dziugaite and Roy bounds*

We notice that this relation was independently explored by Dziugaite and Roy, who worked on the hypothesis that SGD finds good solutions only if they are surrounded by a large volume of good solutions [DR17], if so, the expected error rate of a classifier drawn at random from this volume should match that of the SGD solution. Theorem 4.7 (Preliminary Theorem 2 [McA99]) bounds the expected error of a classifier chosen from a distribution Q in terms of the D_{KL} divergence from a prior P, and if the volume of good solutions is large, and not too far from the mass of P, we obtain a good bound.

[DR17] Dziugaite and Roy, 'Computing Non-vacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data'. URL: http://auai.org/uai2017/proceedings/papers/173.pdf

They use SGD to optimise the PAC-Bayes bound on the error rate of a stochastic neural network, *i.e.* a DNN that represents a stochastic mapping $p(y|x)$. The objective function is the sum of

- the empirical *surrogate*[20] loss averaged over a random pertubation of the SGD solution;
- a generalisation error bound that acts as a regulariser.

Recall Corollary 1:

Corollary 1. *Any learning algorithm that:*

- *assumes a stochastic* $p(y|x)$*;*
- *uses a* D_{KL}*-equivalent loss (for example the cross-entropy loss or the logistic loss);*
- *and a regularisation term that penalises the amount of information of the input stored in the model,*

is learning a minimal sufficient disentangled representation and, in fact, solving the IB learning problem.

From this, we can say that Dziugaite and Roy is solving an instance of the IB learning problem.

Moreover, their objective can be written as [DR17, sec. 6]:

$$\min \mathbb{E}_{W\sim\mathcal{N}} L(W, S) + [w - w_0]^\top \operatorname{diag}(s)[w - w_0] \tag{8.39}$$

where s is the score function. In other words, they are calculating the diagonal of the Fisher Information Matrix as a regularizer.

8.9 EVIDENCE OF THE IB LIMIT IN A HUMAN LEARNED TASK

Zaslavsky et al. [Zas+18] had the sagacious idea of using the IB method to analyse anthropological evidence.

We have already established that intelligent agents, whether artificial or biological, need language to represent a complex environment. Natural languages reflect different solutions to this problem. The current most accepted theory in Anthropology and Linguistics suggest that while languages vary to accommodate language-specific needs (due, for example, to variations in the environment), they evolve into efficient representations [Zas+18]. Although not explicit in [Zas+18], it is evident that the evolution of natural languages can be seen as a learning process for the task of efficient communication by a society.

The paper analyses natural languages in the context of colour naming. It is based on the World colour Survey (WCS), "*a large colour-naming database obtained from informants of mostly unwritten languages spoken in pre industrialised cultures that have had limited contact with modern, industrialised society*" [LB09]. Assuming that each

[20] The *surrogate* loss is the logistic loss, which is differentiable.

[DR17] Dziugaite and Roy, 'Computing Non-vacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data'. URL: http://auai.org/uai2017/proceedings/papers/173.pdf

[Zas+18] Zaslavsky et al., 'Efficient compression in color naming and its evolution'.

[LB09] Lindsey and Brown, 'World Color Survey color naming reveals universal motifs and their within-language diversity'. URL: https://www.pnas.org/content/106/47/19785

colour of WCS corresponds to a specific meaning; it formulates the problem of colour naming in an information-theoretical perspective analogous to the IB problem setting [TPB99]:

[TPB99] Tishby et al., 'The Information Bottleneck Method'.

Target	Intended Meaning	Source	Encoder	Representation	Decoder	Estimated Target
Y	$\xrightarrow[p(y\|x)]{}$	X	$\xrightarrow[q(z\|x)]{}$	Z	$\xrightarrow[q(y\|z)]{}$	$\hat{Y}$
U	$\xrightarrow[m(u)]{}$	M	$\xrightarrow[q(w\|m)]{}$	W	$\xrightarrow[\hat{m}_w(u)]{}$	$\hat{M}$
Universe (colour)	Intention	Source	Language (naming distribution)	Word	Language	Interpretation

FIGURE 8.5: Adapted from [Zas+18]

With that formulation, it is possible to calculate the IB limits and analyse the different languages colour-naming solutions in this framework:

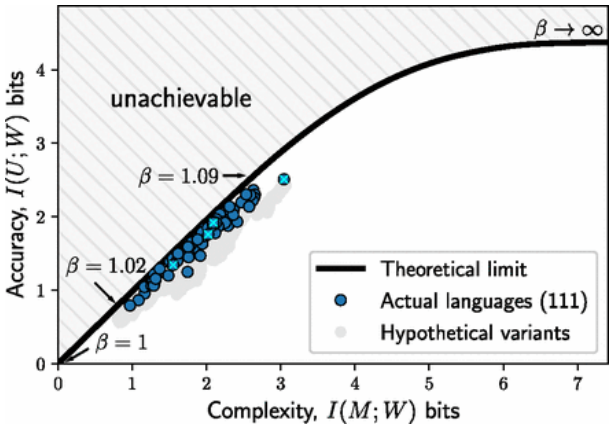


FIGURE 8.6: Different languages (blue circles) achieve near-optimal compression (IB curve in black) [Zas+18]. Reproduced with permission.

In Figure 8.6, it is possible to see evidence that languages efficiently compress ideas into words by optimising the tradeoff between complexity and accuracy of the lexicon according to the IB principle.

This analysis corroborates the current theory on human language evolution. Furthermore, the drive for information-theoretical efficiency explains why human languages categorise colour as they do and may also apply to learning in general.

The hypothesis is that languages evolve to become more efficient in a tradeoff between conciseness (complexity, generalisation) and precision. The prediction capability is just an expected consequence of an efficient representation of meaning. The conciseness of the representation of knowledge, given an acceptable error margin, is a proxy of the agent's intelligence. The IB limit is an epistemic limit that is valid for machines, humans and aliens.

8.10 CONCLUDING REMARKS

This chapter presented the IBT as a general representation learning theory (not specific to Deep Learning). The bulk of this chapter is based on works by Stefano Soatto and Alessandro Achille and their prolific research group [AS18a; Ach19; AS19; AS18b]. 'Emergence of Invariance and Disentangling in Deep Representations', in particular, has been one of the biggest influences in this dissertation. It was presented in the same workshop[21] where Tishby presented IBT for the first time.

[AS18a] accomplishments in this chapter were threefold:

1. It explained the emergence of invariance (generalisation) and disentanglement in the proposed learning setting.

2. It addressed one of the weaknesses of IBT: the confusion about past and future data. MLT provides rigorous guarantees for future performance (test time) based on the past data (training time). Conversely, several of the initial IBT papers were not clear with what is happening during training and how it is different in test time.

3. It showed the crucial role of noise and how it can be controlled in favour of generalisation.

4. It demonstrated that the information in the weights, despite being difficult to measure, can be bounded by the Fisher information Matrix:

$$\mathrm{I}[Z;X] \leqslant \mathrm{I}[W;D] \leqslant \log|\mathrm{F}(\mathrm{W}^*)| \leqslant \log|\mathrm{F}(\mathrm{W})|$$

Noteworthy, the Deep Learning setting does not seem to correspond to the conditions of Corollary 1, as [Zha+16] has shown that Deep Learning converges even in the absence of a regulariser in the loss function.

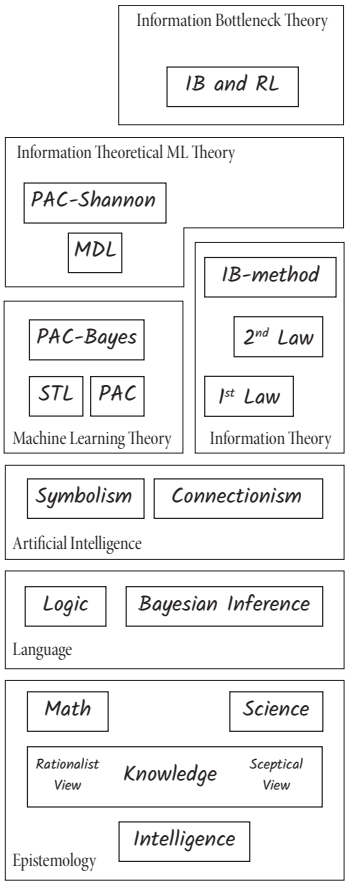


FIGURE 8.7: In this chapter we derived a learning formulation from a set of desired properties for representations that relates to the IB principle.

8.10.1 *Assumptions*

1. MLT assumptions

2. Information is what changes belief.

3. IBT for Representation Learning assumptions:

   i. The random variables X, Y and Z are **discrete**;

[21] Deep Learning: Theory, Algorithms, and Applications. Berlin, June 2017 http://doc.ml.tu-berlin.de/dlworkshop2017

[AS18a] Achille and Soatto, 'Emergence of Invariance and Disentangling in Deep Representations'.

ii. $Y \to X \to Z$ form a Markov-chain;

iii. $\mathbb{A}_X$, $\mathbb{A}_Y$ and $\mathbb{A}_Z$ are **finite sets**;

iv. **No assumption on** $\mathrm{D} = \mathrm{P}(X, Y)$.

v. $\mathrm{D} = \mathrm{P}(X, Y)$ **is unknown at the training stage**.

vi. $\mathrm{D} = \mathrm{P}(X, Y)$ **is fixed:** the ordering of examples in the sample is irrelevant.

vii. X is is i.i.d. sampled.

viii. The encoder and the decoder are **stochastic** mappings.

ix. the loss function is in the form of a IB Lagrangian (Corollary 1), *i.e.* $\mathcal{L}(W) = \mathrm{H}_{p,q}[D|W] + \beta^{-1}\mathrm{I}[D; W]$ has a regulariser term that penalises the memorisation of the dataset.

We took the liberty to add the assumption that constrains the problem to finite alphabets (discrete random variables). Unfortunately, with few exceptions, the literature on IBT does not underscore this constraint nor, alternatively, demonstrate why one can use differential entropy.[22]

[22] For example, Alemi et al. use differential entropy but do not address the fact that the IB Principle restrain itself to discrete random variables [Ale+16].

9

The Information Bottleneck and Deep Learning

'Great claims require great evidence.'

—Carl Sagan

This chapter presents IBT for Deep Learning, the context where all IBT papers focus. All previous chapters brought concepts needed to understand this chapter. In chronological order, the research from which Chapter 7 is based was published almost 20 years earlier than the contents presented in Chapter 8, which were published more or less simultaneously as the contents of this chapter.

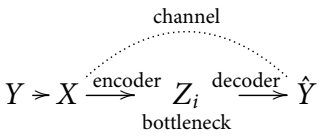
9.1 DEEP LEARNING IN THE IBT PERSPECTIVE

In MLT, the analysis of learning algorithms is based on a hypothesis space. This choice may have biased the Deep Learning community focus on architectures. For many, Deep Learning (DL) and Deep Neural Networks (DNNs) are interchangeable names.

The IBT perspective has a holistic view of Deep Learning (DL) where each of its components has a role.

9.1.1 *Deep Neural Network in IBT*

IBT assumes that DNN layers are random variables that form a Markov chain from the target variable to the prediction. Each layer is a representation Z_i of the input at a different "resolution"/abstraction (Section 7.4). These representations act like bottlenecks in the input-output channel. Thus, each bottleneck defines a unique encoder/decoder scheme.



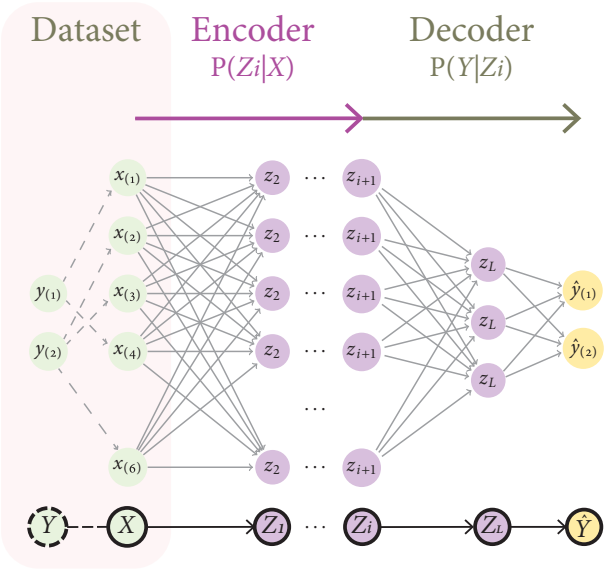


FIGURE 9.1: A Deep Neural Network as a Markov-chain in the Information Bottleneck perspective.

9.1.2 *SGD in IBT*

One of the most contentious topics in IBT is the assumption that $q(z|x)$ and $q(y|z)$ are stochastic. Noise plays a very important role in training [HVC93; AS18a; KSW15]. In IBT, noise reduces capacity and, therefore, the size of the typical hypothesis space (as it will be shown in Section 9.5.2).

Counter-intuitively, Chaudhari and Soatto [CS18] proves (with theory and extensive empirical evidence) that SGD performs variational inference for a different loss than the one used to compute the gradients and that this loss has a regulariser term that is equivalent to the information bottleneck principle (Corollary 1).

[HVC93] Hinton and Van Camp, 'Keeping the neural networks simple by minimizing the description length of the weights'.

[AS18a] Achille and Soatto, 'Emergence of Invariance and Disentangling in Deep Representations'.

[KSW15] Kingma et al., 'Variational Dropout and the Local Reparameterization Trick'. URL: https://proceedings.neurips.cc/paper/2015/file/bc7316929fe1545bf0b98d114ee3ecb8-Paper.pdf

[CS18] Chaudhari and Soatto, 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'.

9.1.3 *Loss function in IBT*

The IB Principle (Chapter 7) provides compelling grounds for the use of the Kullback-Leibler divergence (D_{KL}) as the canonical loss function. It is equivalent by a constant to the cross-entropy loss, which became ubiquitous in DL (as shown in Section 8.4).

9.2 LITERATURE

We are using the name Information Bottleneck Theory (IBT) as an "umbrella" to designate the work that relates to our selected literature (Appendix A). Frankly, the designation has not been adopted consistently. Nonetheless, we can identify three kinds of literature:

1. IB-based analysis of Deep Learning

2. IB-Deep Learning applications
3. IB-based theory of Deep Learning

We will detail each kind of literature in the following sections.

9.3 IB-BASED ANALYSIS OF DEEP LEARNING

9.3.1 *Opening the black-box: the information plane*

One of the critiques on current MLT is on its choice of treating the models as black-boxes (Section 4.8.1). This choice allows MLT to be more general, independent of the class of hypothesis. At the same time, the current theory provides little guidance for what happens during training, letting the community figure out many possible competing explanations.

There is nothing wrong with the choice. It may be an advantage in most cases. But in the case of Deep Learning, where there are still many phenomena with no clear winner explanation and where there is a growing demand for understanding why DNNs make this or that choice, a different choice may help.

This is what motivated Shwartz-Ziv and Tishby [ST17], according to Tishby himself [Tis20]. Shwartz-Ziv and Tishby propose using the mutual information between the activations in different layers and the input. Despite being a measure difficult to calculate, it has the potential of "opening the black-box", *i.e.* it allows in Tishby's words to see training with an "X-Ray" [Tis20]

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

[Tis20] Tishby, *The Information Bottleneck View of Deep Learning: Why do we need it?*. URL: https://youtu.be/utvIaZ6wYuw

9.3.2 *Information Plane and Deep Learning*

Shwartz-Ziv and Tishby hypothesis was that the information-plane (Section 7.3.1) could be their "X-Ray" [ST17]. To overcome the difficulty of calculating the mutual information[1], they created a synthetic

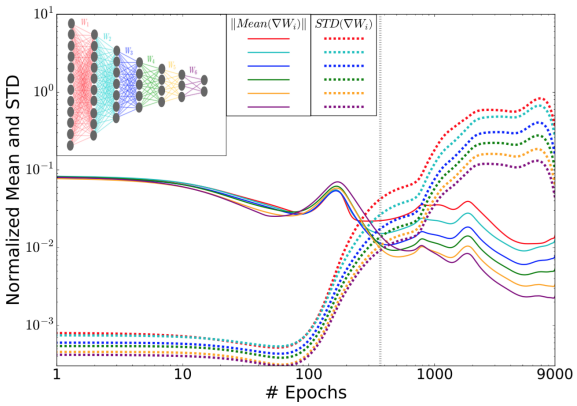


FIGURE 9.2: The plot of the norm of mean and standard deviation of the layers weight gradients as a function of training epochs shows two distinct phases. (Reproduced from: [ST17])

[1] By the time of their paper, there was no known algorithm to calculate the mutual information for discrete X and Y with large state spaces or non-Gaussian continuous joint distribution. Eventually, [Ale+16] and [AS18b] independently invented such algorithm using the variational formulation of the IB Lagrangian (Section 9.4.1).

dataset for which they knew in advance the usually unknown distribution $P(X, Y)$, added a noisy layer to guarantee the stochastic mapping[2] and calculated the mutual information during training with a binning strategy.

The result was visually appealing (Figure 9.2). It clearly shows a *phase transition* during training (Figure 9.3).

[2] A kind of *test-time augmentation*, a common practice in ML that injects noise to a test input, by generating transformed versions of it with slightly different $\hat{y}_i$ in different runs of the model for the same x_i., and combines the predictions of these versions.

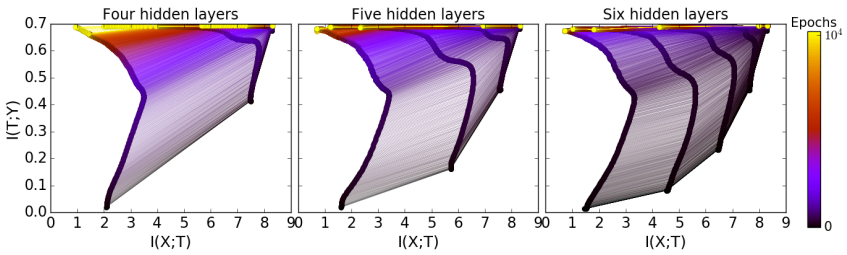


FIGURE 9.3: The information planes for different architectures show the amount of information in the layers during SGD optimisations. It is possible to see in all images that there is a fast *fitting* phase where the model rapidly obtain information about the target variable by memorising the input. Most of the training time, however, is spent in a *compression* phase where the model tries to *forget* as much as possible of the input while keeping the relevant information of the target variable (Reproduced from: [ST17]).

9.3.3 *IBT's main thesis*

In Tishby's words, IBT main thesis can be summed up as *"learning is forgetting"*[3]. More specifically, deep learning has two distinct training phases:

Fitting Phase: When the DNN rapidly (in terms of epochs) overfits to the training data;

Compression Phase: When the DNN compresses the amount of information, *forgetting* as much it can about the input, while keeping the relevant information about the target;

[3] An observation that was made longer before by Chaitin in [Chao6].

In statistical mechanics, phase transitions relate to abrupt changes in the properties of a system at the macroscopic level, in the same way as seen in Figure 9.2. With that in mind, Shwartz-Ziv and Tishby claim that the compression phase can be described by Focker-Plank diffusion equations from Physics. This was indeed later corrobored by Chaudhari and Soatto [CS18; Cha+19a], but Shwartz-Ziv and Tishby failed to support the claim that DNNs can be seen as physical systems.

9.3.4 *Criticism to IBT's main thesis*

Shortly after its publication, Shwartz-Ziv and Tishby [ST17] were challenged by Saxe et al. [Sax+18], who claimed that they could not replicate the experiment and argued that the binning procedure to estimate mutual information was inexact. Due to the fact that the

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

[Sax+18] Saxe et al., 'On the Information Bottleneck Theory of Deep Learning'.

activation function can be an invertible transformation (deterministic mapping) of the input, by reparametrisation invariance (RI), the true mutual information between $\mathrm{I}[X;Z_\mathrm{L}]$ is provably infinite for continuous distributions and constant (*i.e.* equal to $\mathrm{H}[X]$) for discrete ones. They also point out that a user-selected binning strategy leads to arbitrary values of mutual information in the plotted results. Overall, Saxe et al. refute Shwartz-Ziv and Tishby results.

Other authors followed their reservations in different degrees: Goldfeld et al. [Gol+19] agree that Shwartz-Ziv and Tishby's $\mathrm{I}[X;Z_\mathrm{L}]$ estimates do not directly measure compression of the true mutual information. Chelombiev et al. [CHO19] explore several estimation schemes and were able to measure compression but with several caveats.

This relates to one of the weaknesses of IBT: lack of rigour that even Tishby admits [Tis20]: *'I would not call [IBT] a proven rigorous theory.'* If, on one hand, their spectacular claims have driven much interest to the subject, on the other it generated an equivalent dose of suspicion and scrutiny. As Carl Sagan once said, *'great claims require great evidence'*.

Some IBT papers failed to point out that the IB Principle is ill-posed for deterministic functions. Therefore, there is a missing argument of why and in which conditions we can see the function of the activations as a stochastic mapping. Bayesian interpretations may justify parameter noise, but activation noise has no such theoretical ground.

[Gol+19] Goldfeld et al., *Estimating Information Flow in DNNs*.
URL: https://openreview.net/forum?id=HkxOoiAcYX

[CHO19] Chelombiev et al., 'Adaptive Estimators Show Information Compression in Deep Neural Networks'.

[Tis20] Tishby, *The Information Bottleneck View of Deep Learning: Why do we need it?*.
URL: https://youtu.be/utvIaZ6wYuw

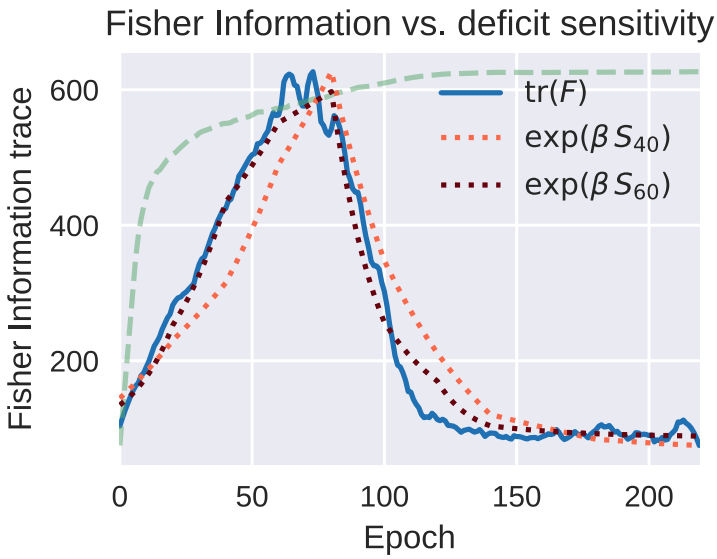


FIGURE 9.4: Information in the weights drop abruptly and the model keeps improving in test-time. The amount of information in the weights measured by Achille et al. corroborates with Tishby's thesis (Section 9.3.3). Reproduced from [ARS17].

In time, Tishby's conjectures and intuitions were corroborated by others findings. In special, Stefano Soatto and his research group not only observed the fitting-compression phases (Figure 9.4) but also proved a crucial missing step: information in the activations of future data is bounded by the information in the weights during training, where stochasticity can be explained [AS18b] (Section 8.5).

[AS18b] Achille and Soatto, 'Information Dropout: Learning Optimal Representations Through Noisy Computation'.

Besides, to prove this theoretical result, they created a variational method (equivalent to Deep Variational Information Bottleneck (DVIB), Section 9.4.1) for estimating mutual information using Deep Learning[4], obtaining more accurate mutual information measurements. In another venue, Chaudhari et al. corroborated the statistical mechanics' intuition for the behaviour of SGD with experimental results [Cha+19a].

[4] Deep Learning helping to understand deep learning.

[Cha+19a] Chaudhari et al., 'Entropy-SGD: Biasing gradient descent into wide valleys'.

9.4 IB-BASED DEEP LEARNING APPLICATIONS: TRAINING AND ALGORITHMS

9.4.1 *Deep Variational Information Bottleneck (DVIB)*

A common criticism on IBT was related to difficulties in calculating mutual information (Section 9.3.4). DVIB not only describes a loss metric that takes advantage of IB properties but also defines state-of-the-art approximations of $\mathrm{I}[Z;X]$ and $\mathrm{I}[Z;Y]$ [Ale+16].

[Ale+16] Alemi et al., 'Deep variational information bottleneck'.

Tishby and Zaslavsky already envisioned using the IB to train DNNs [TZ15a]. Tishby, however, wanted IBT to be seen as IB-based analysis tool. Subsequently, he believed that IB-based applications "miss the point" that IBT works even if you do not know anything about the IB [Tis20].

[TZ15a] Tishby and Zaslavsky, 'Deep learning and the information bottleneck principle'.

[Tis20] Tishby, *The Information Bottleneck View of Deep Learning: Why do we need it?*. URL: https://youtu.be/utvIaZ6wYuw

Still, Alemi et al. considered the idea of using the IB in training appealing as it defines a good representation in terms of the trade-off between a conciseness and predictive power. They noticed, however, that the main drawback in using it in practice was that calculating the mutual information is challenging. The proposed method solves this drawback.

Curiously, the proposed method is equivalent to the variational inference presented in 'Information Dropout: Learning Optimal Representations Through Noisy Computation' [AS18b]. This similarity was noticed by the authors themselves that despite not citing [AS18b] in the first version, cited it in subsequent versions. Despite of the concurrent idea development, the organisation and clear focus made DVIB the prefered reference for using the IB objective to estimate information measures.

DVIB became essential to evaluate the claims of Tishby that, during training, DNNs experience two distinct phases, fit and compression.

Deep Variational Information Bottleneck Method

Let us formulate a variational IB:

$$\theta^* = \arg\max_{\theta} I[Z;Y|\theta] \text{ s. t. } I[X;Z|\theta] \leqslant I_c. \tag{9.1}$$

$$R_{IB}(\theta) = \underbrace{I[Z;Y|\theta]}_{(A)} - \underbrace{\beta I[Z;X|\theta]}_{(B)} \tag{9.2}$$

where θ is the set of parameters of the network. This IB Lagrangian formulation has two parts (A and B). Notice that

$$I[Z;Y] = H_p[Y] - H_p[Y|Z], \tag{A}$$

where $p(y|x)$ and $p(x)$ are unknown, which makes part A intractable. Let $q(y|z)$ be the variational approximation, our decoder, which will be another DNN with its own parameters, which is tractable.

$$D_{KL}(p\|q) \geqslant 0 \rightarrow H_p \geqslant H_q \tag{9.3}$$

$$\therefore I[Z;Y] \geqslant \cancel{H_p[Y]}^{\text{constant}} - H_q[Y|Z] \tag{9.4}$$

$$\geqslant -H_q[Y|Z] = \sum_{x,y,z} p(y|z)p(z|x)p(x)\log q(y|z). \tag{9.5}$$

And now part B:

$$I[Z;X] = D_{KL}(p(z|x)//p(z)). \tag{B}$$

But p(z) might be difficult to calculate. So, let r(z) be a variational approximation of this marginal. Since $D_{KL}(p//r) \geqslant 0$,

$$I[Z;X] \leqslant D_{KL}(p(z|x)//r(z)) \tag{9.6}$$

$$\leqslant \sum_{x,y,z} p(y|z)p(z|x)p(x)\log q(y|z). \therefore \tag{9.7}$$

$$I[Z;Y] - \beta I[Z;X] \geqslant \sum_{x,y,z} p(y|z)p(z|x)p(x)\log q(y|z) - \beta \sum_{x,y,z} p(y|z)p(z|x)p(x)\log q(y|z) = L. \tag{9.8}$$

Approximating L empirically:

$$L \approx \frac{1}{N}\sum_{1}^{N}\left[\sum p(z|x_n)\log q(y_n|z) - \beta p(z|x_n)\log\frac{p(z|x_n)}{r(z)}\right]. \tag{9.9}$$

Which can be solved using the reparametrisation trick [KSW15].

[KSW15] Kingma et al., 'Variational Dropout and the Local Reparameterization Trick'. URL: https://proceedings.neurips.cc/paper/2015/file/bc7316929fe1545bf0b98d114ee3ecb8-Paper.pdf

9.4.2 *Information Dropout*

'Information Dropout: Learning Optimal Representations Through Noisy Computation' establishes links between different and seemly unrelated research topics as dropout, variational auto-encoders and optimal representations through the IB principle. Its theoretical development is not being the paper focus, is its most important contribution to IBT. In this sense, the method that names the paper is just a way to empirically support their interesting theoretical claims (Chapter 8).

Nevertheless, the technique is a generalisation of the well-known Dropout method. Chaudhari and Soatto theoretically suggest that noise intrinsic to the architecture (dimensionality reductions, dropout, small mini-batches, *etc.*) is better for generalisation than noise in the dataset [CS18]. In this sense, there are research opportunities in exploring Information Dropout and other forms of controlling the information in the weights with the injection of noise. In areas like NLP, where data-augmentation is challenging, Information Dropout may play an important role.

[CS18] Chaudhari and Soatto, 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'.

The emergent properties of representations, the generalisation of dropout and the connection to variational autoencoders are surprising results that should be of interest to researchers in representation learning (Section 8.7).

9.4.3 *Transferability metrics*

To this day, transferability is measured experimentally or inferred subjectively by experts according to tasks "proximity" [Zam+18]. 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.

[Zam+18] Zamir et al., 'Taskonomy: Disentangling task transfer learning'.

This illustrates the importance of building such a topology. In other words, we want 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?

Intuitively, the complexity of a learning task is related to its best expected out-of-sample error.

Given a fixed architecture, the amount of information in the weight measures how much "memorisation" was used to fit the model. High information in the weights suggests more *difficult* tasks. The Fisher Information Matrix (FIM) measures the resilience of the loss due to perturbation in the weights (Figure 9.5). If a weight accepts more noise (*i.e.* it can be perturbed without a significant change in the model error), it is less important, and there is no need to "memorise" it. Also, this amount of noise has a direct correspondence to generalisation (Section 9.5.1). Using this intuition, Achille et al. uses the diagonal of the FIM as an embedding that represents the task itself. Since the FIM can be too noisy when trained from a few examples, the diagonal of the FIM is used as it is considered a more simple and robust representation [Ach+19].

[Ach+19] Achille et al., 'Task2Vec: Task Embedding for Meta-Learning'.

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 $p(y|z)$, which usually can be done efficiently, and then computing the FIM for the feature extractor parameters.

Transferability (or fine-tuning 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 initialisation, t_0, and the performance of a solution to t_a fine-tuned for t_b:

$$\mathrm{D}_{\text{f-t}}\left(t_a \to t_b\right) = \frac{\mathbb{E}\left[\ell_{a\to b}\right] - \mathbb{E}\left[\ell_b\right]}{\mathbb{E}\left[\ell_b\right]}, \tag{9.10}$$

where expectations are taken over all training, ℓ_b is the final test accuracy obtained by training task b from initialisation, 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 is due to its high complexity.

9.5 IB-BASED DEEP LEARNING LEARNING THEORY

In Section 8.10, we concluded with a seemly missing step of IBT in the context of Deep Learning: the fact that Corollary 1 requires an information-limiting regulariser in the loss function, which is not explicitly present in many DL models that converge. In this chapter, however, we presented the work of Chaudhari and Soatto who showed

that even if there is no explicit regulariser, the use of SGD guarantees it is implicitly there.

Another assumption of IBT learning is that the task is a stochastic mapping between the input and output. In the context of Deep Learning, with its large datasets, this is hardly a limitation.

An important theoretical discussion specific to Deep Learning that has not been addressed yet is about the role of layers. This will be the subject of Section 9.5.2.

9.5.1 *A new narrative*

[GBC16] Goodfellow et al., *Deep Learning*.

According to Goodfellow et al. [GBC16], Deep Learning success is ascribed to several pleasant features for which our current understanding is largely empirical. Here, we use Information Bottleneck Theory's (IBT) most crucial strength, its narrative, to give theoretical ground to some DL phenomena.

Generalisation power despite a huge number of parameters As we have already shown in Section 6.2, the complexity of a task relates to the amount of information needed to describe it. In this sense, even if the network has a nominal capacity that relates to the parameters, its effective capacity is the mutual information $\mathrm{I}[X;Y]$ (or $\mathrm{I}[X;Y|W]$). This interpretation of complexity does not invalidate the complexity-performance trade-off in MLT.

Generalisation despite expressiveness–overfitting For high-capacity models, generalisation has to do more with *overfitting* than *underfitting*. We have shown that the loss function that *emerges* from a definition of good representations (Section 8.2), has an implicit overfitting term that can be neutralised (Section 8.4).

To neutralise the effect of overfitting, the loss needs a regulariser term that penalises the model for keeping information about the training dataset. Even if this term is not explicitly added to the loss function, Chaudhari and Soatto shows that it is implicitly there [CS18].

[CS18] Chaudhari and Soatto, 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'.

Deep Learning bias for disentangled representations This happens because the implicit regulariser term in SGD is in a form that is equivalent to the assumption that the representation has zero multi-information, *i.e.* no correlation between its components. This property relates to disentanglement.[5]

[5] In IBT, disentanglement is defined as this property.

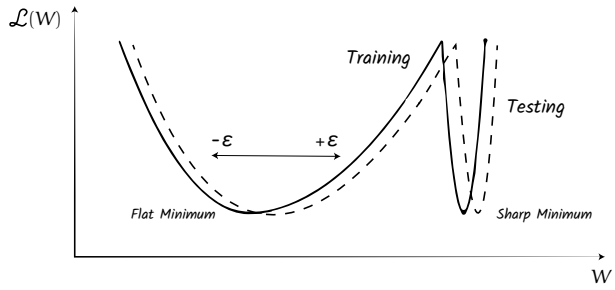


FIGURE 9.5: Information in the weights explain the preference of SGD for flat minima. In regions of flat minima, the effect of noise (dashed line) is minimal.

SCARCITY OF BAD MINIMA ENCOUNTERED BY SGD OPTIMISATION It is a known fact that SGD optimisation tends to find "flat minima", regions in the weight space where small perturbations in the value of the weight leads to similar small error (Figure 9.5) [HS97; Mac92]. Mackay already explained, via Bayesian inference, that this relates to small information in the weights (amount of information affects the curvature of the space) [Mac92].

[HS97] Hochreiter and Schmidhuber, 'Flat minima'.

[Mac92] Mackay, 'The Evidence Framework Applied to Classification Networks'.

This explanation is consistent with IBT perspective. As we have already shown, the information in the weights is bounded by the Fisher information in the weights that measures the curvature of the weight space. Another interesting implication of this information interpretation is that due to the AEP all local minima have approximately the same chance of being found in the weights typical space.[6]

[6] We use this property to show that layers help to find local minima, Section 9.5.2.

CRITICAL-LEARNING PERIODS Critical-learning periods are time windows of early development during which sensory deficits can lead to permanent skill impairment. These are well-documented phenomena in humans, and other animals [Wie82]. Surprisingly, Achille et al. show that DNNss exhibit such critical periods as well [ARS17] . This finding questions the assumption that the order in which a model experiences evidence does not affect learning.

[Wie82] Wiesel, 'Postnatal Development of the Visual Cortex and the Influence of Environment'.

[ARS17] Achille et al., *Critical Learning Periods in Deep Neural Networks.*

In their experiments, Achille et al. used the Fisher Information Matrix (FIM) of the weights to measure information in the network. They caused sensory deficits by blurring input images and noticed that such deficits cause the information in the weights to grow and remain higher even after they are removed. This deficit may be attributed to forcing the network to memorise the labels.

The IBT explanation for such phenomena is due to the training phase transition [ST17]. In the first phase, the network moves towards high-curvature regions of the loss landscape, while in the second phase, the curvature decreases, and the network eventually converges to a flat minimum.

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

Analysing Figure 9.6, we can see that networks more affected by

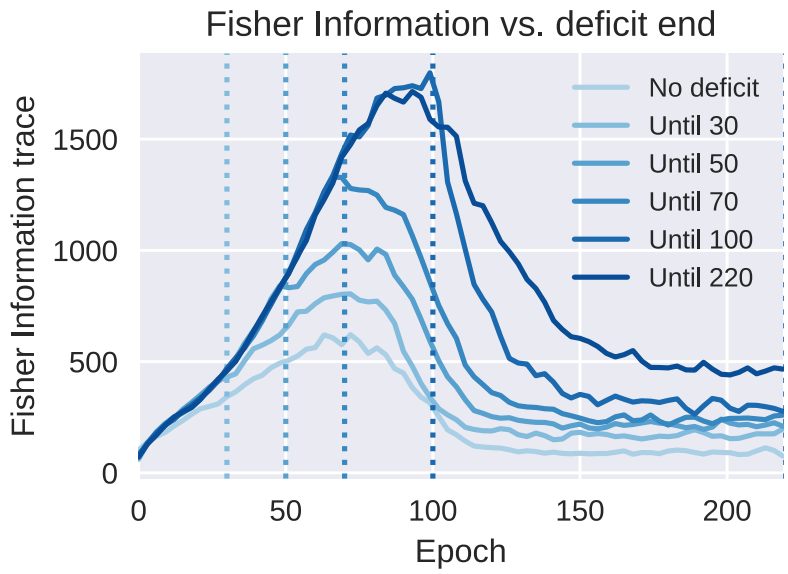


FIGURE 9.6: Information in the weights grow in the event of a sensory deficit and remains higher even after the deficit is removed.

the deficit converge to relative sharper minima.

During the first phase, with a sensory deficit, the network is obliged to cross regions of high curvature in the loss geometry in order to achieve a certain performance before eventually entering a flatter region of the loss surface and ending up trapped in the higher curvature region.

THE ROLE OF LAYERS IN DEEP LEARNING This will be explained in a section of its own (Section 9.5.2)

9.5.2 *The role of layers in deep learning*

Why do we need multiple layers in a neural network? This question is fundamental in Deep Learning, and still, there is no definitive answer. A feedforward network with a single layer can represent any function [GBC16]. Also, Leshno et al. [Les+93] (as cited by [GBC16]) demonstrated that shallow networks with rectified linear units as activation functions have universal approximation properties. When confronted with these facts, the usual answer for the need for depth is that these results require an infeasible large layer or do not address efficiency. Another common answer is that layers provide levels of abstraction and a paramount composability property, *i.e.* stacking layers allow a network to represent functions of increasing complexity [GBC16]. These answers seem correct but, at the same time, somewhat qualitative and vague.

[GBC16] Goodfellow et al., *Deep Learning*.

[Les+93] Leshno et al., 'Multilayer feedforward networks with a nonpolynomial activation function can approximate any function'. URL: https://www.sciencedirect.com/science/article/pii/S0893608005801315

This section will try to advance the discussion by answering the need for depth in Neural Networks with an Information Bottleneck perspective. Shwartz-Ziv and Tishby has provided an explanation based in Physics [ST17]. Here we will not use such correspondence.

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

We have already established that a DNN optimised with SGD solves an IB problem. In this view, the body of the network is an encoder that compresses the input X into a representation Z. In the IBT perspective, training a DNN is finding the encoder that minimises $\mathrm{I}[Z;X]$, while keeping $\mathrm{I}[Z;Y]$:

$$\mathrm{Q}(Z|X) \mid \mathrm{Q} = \arg\min \mathrm{I}[Z;X] \text{s.t.} \mathrm{I}[Z;Y] \geqslant \mathrm{I}_{\mathrm{Y}}$$

From [AS18a]:

[AS18a] Achille and Soatto, 'Emergence of Invariance and Disentangling in Deep Representations'.

Corollary 2 (Bottlenecks promote invariance)**.** *Assume a Markov chain of layers:*

$$X \to Z_1 \to Z_2,$$

and that there is a bottleneck between Z_1 and Z_2 (for example, if $dim(Z_1) > dim(Z_2)$ or noise has been added between to the channel $Z_1 \to Z_2$ via dropout[7]). Then, if Z_2 is sufficient, it is more invariant to nuisances than Z_1 (see Section 8.2.1).

[7] Dimensionality reduction can be seen as a form of noise.

Corollary 3 (Stacking increases invariance)**.** *Assume a Markov chain of layers:*

$$X \to Z_1 \to Z_2 \to \cdots \to Z_{\mathrm{L}},$$

and that Z_{L} is sufficient of X w.r.t. Y. Then, by DPI:

$$\mathrm{I}[Z_{\mathrm{L}};X] \leqslant \mathrm{I}[Z_i;X], \forall i \in \{1 \cdots \mathrm{L}-1\},$$

therefore Z_{L} is more insensitive to nuisances than all preceding layers and generalises better.

In other words, Achille and Soatto argue that **stacking layers improve generalisation** [AS18a]. A problem with this argument is that it only shows that in the multi-layered scenario, the last layer is more compressed and invariant to nuisances. It does not contradict that a single-layered network could achieve the same level of compressibility of the input as the last layer in the multi-layered scenario.

Besides, according to Achille and Soatto, the above corollary does not simply imply that the more layers, the merrier, as there is

the assumption that one has successfully trained the network (Z_L is sufficient). For Achille and Soatto, a successfully trained network becomes increasingly difficult as the network grows. The increase in difficulty seems straightforward because stacking layers increase the number of computations per batch.

By pure logic, it is evident that the complexity of an algorithm that searches for the best possible hypothesis will depend on the size of the hypothesis space. Counter-intuitively, however, we claim that **stacking layers decreases the complexity of the training in practice**, as they reduce the "typical" hypothesis space size. We argue that stacking layers increases the complexity of the learning algorithm by a constant while exponentially decreasing the size of the "typical" hypothesis space.

To explain how stacking layers reduces the "typical" hypothesis space size, let us start with the case of a single layer DNN. We need to find:

$$Q_1^* = Q_1(Z_1|X) \mid Q_1^* = \arg\min I[Z;X] \text{s.t.} I[Z;Y] \geqslant I_Y \tag{9.11}$$

What is the "typical" hypothesis space in this case? First, let $\mathbb{A}_{Q_1}$ be the hypothesis space of all functions that encode $x \in |\mathbb{A}_X|$ possible inputs into $z \in |\mathbb{A}_Z|$ possible outcomes. Nevertheless, from Shannon's Noisy Channel Theorem (Section 5.7) there are only $2^{H[X]}$ typical x_i and for each typical x_i there is only $2^{H[Z|X]}$ "typical" outcomes (see Section 5.5.6), therefore, the number of decodable encodings (one to one mappings) is:

$$\forall q \sim \mathbb{A}_{Q_1}, Pr(q \in \mathbb{A}_{Q_1}^{\delta}) = 1 - \delta, \delta \to 0 \tag{9.12}$$

$$|\mathbb{A}_{Q_1}^{\delta}| = \frac{2^{H[X]}}{2^{H[Z_1|X]}} = 2^{I[Z;X]}. \tag{9.13}$$

where δ can be arbitrarily small.

Now, let us compare with the case of L layers. Each layer acts as an encoder $q_i : Z_{i-1} \to Z_i$. The typical hypothesis space of q_i depends on the cardinality of the possible values of z_i, which will depend on the size of the typical hypothesis space of q_{i-1} (a one-to-one map):

$$|\mathbb{A}_{Q_i}^{\delta}| = \frac{|\mathbb{A}_{Q_{i-1}}^{\delta}|}{2^{H[Z_i|Z_{i-1}]}}. \tag{9.14}$$

$$\mathrm{Q_L} = \mathrm{Q}(Z_\mathrm{L}|Z_{\mathrm{L}-1}) \circ \mathrm{Q}(Z_{\mathrm{L}-1}|Z_{\mathrm{L}-2}) \circ \cdots \circ \mathrm{Q}(Z_1|X) \tag{9.15}$$

$$\therefore \tag{9.16}$$

$$|\mathbb{A}^{\delta}_{\mathrm{Q_L}}| = \frac{2^{\mathrm{H}[X]}}{2^{(\mathrm{H}[Z_1|X]+\mathrm{H}[Z_2|Z_1]+\cdots+\mathrm{H}[Z_{\mathrm{L}-1}|Z_\mathrm{L}])}} \tag{9.17}$$

$$= 2^{\mathrm{I}[Z_1|X]-(\mathrm{H}[Z_2|Z_1]+\cdots+\mathrm{H}[Z_{\mathrm{L}-1}|Z_\mathrm{L}])} \tag{9.18}$$

We still need to show that $\mathrm{H}[Z_i|\mathrm{Z}_{i-1}] \neq 0, i \in 1 \ldots \ell, X = \mathrm{Z}_0$, *i.e.* $\mathrm{I}[\mathrm{X};\mathrm{Z}_{\mathrm{L}=1}]$ in Q_1 is greater than $\mathrm{I}[\mathrm{X};\mathrm{Z}_\mathrm{L}]$ in $\mathrm{Q_L}$ even if $dim(Z_\mathrm{L})$ is the same in both cases. Besides the dimensionality reduction, even if we disconsider the fact that with more nodes we can explicitly add more noise via dropout or other techinique, there are still the implicity increase of quantisation noise caused by the added layers (with dimensionality reduction). Therefore, $\mathrm{H}[Z_i|\mathrm{Z}_{i-1}] > 0$.

$$\mathrm{H}[Z_i|\mathrm{Z}_{i-1}] > 0 \rightarrow (|\mathbb{A}_{\mathrm{Q_L}}| < |\mathbb{A}_{\mathrm{Q}_1}|)\,. \tag{9.19}$$

This is a direct consequence of corollary 3, $(\mathrm{I}[Z_\mathrm{L}|X] < \mathrm{I}[Z_1|X]) \rightarrow (|\mathbb{A}_{\mathrm{Q_L}}| < |\mathbb{A}_{\mathrm{Q}_1}|)$.

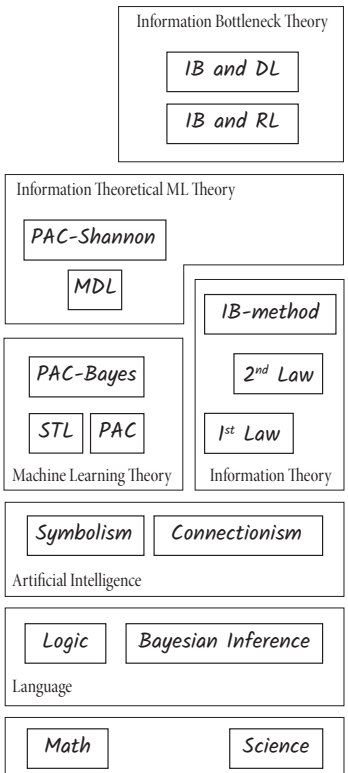
For every bit in $(\mathrm{H}[Z_2|Z_1]+\cdots+\mathrm{H}[Z_{\mathrm{L}-1}|Z_\mathrm{L}])$, the typical hypothesis space is divided by 2. Thus, not only $|\mathbb{A}_{\mathrm{Q_L}}| < |\mathbb{A}_{\mathrm{Q}_1}|$, but $|\mathbb{A}_{\mathrm{Q_L}}| \ll |\mathbb{A}_{\mathrm{Q}_1}|$ (exponentially smaller).

9.6 CONCLUDING REMARKS

This chapter presented the IBT for Deep Learning, showing that it was initially envisioned as an analysis tool to comprehend what happens during training. We also explained why it was received with criticism.

Most of the questions in regards to the lack of rigour were already previously addressed in Chapter 8. In this chapter, we closed the last missing step by showing that even in the absence of an explicit regulariser in the loss function, it is implicitly added by SGD. The acknowledgement of two distinct phases during training may lead to the development of phase-specific training strategies.

Moreover, we demonstrated the power of IBT narrative by giving coherent explanations for several Deep Learning phenomena. For that we did not increase our list of assumptions.



9.6.1 *Assumptions*

1. MLT assumptions:

    a) $\mathrm{D} = \mathrm{P}(X, Y)$ **is unknown at the training stage**.

   b) $\mathrm{D} = \mathrm{P}(X, Y)$ **is fixed:** the ordering of examples in the sample is irrelevant.
   c) X is is i.i.d. sampled.

2. Information is what changes belief.

3. IBT for Representation Learning assumptions:

   i. $\mathrm{D} = \mathrm{P}(X, Y) = \mathrm{P}(Y|X)\mathrm{P}(X)$, where $\mathrm{P}(Y|X)$ is a stochastic mapping.
   ii. The random variables X, Y and W are **discrete**;
   iii. $Y \to X \to W$ form a Markov-chain during training;
   iv. $\mathbb{A}_X$, $\mathbb{A}_Y$ and $\mathbb{A}_W$ are **finite sets**;

10

Conclusion

Our goal was to investigate to what extent the emergent Information Bottleneck Theory (IBT) can help understand generalisation and other Deep Learning Phenomena. In this chapter, we summarise our findings.

10.1 GENERALISATION IN IBT

Foremost, we presented information in the weights as a measure of complexity, a measure with no apparent paradox between generalisation and the number of parameters (Chapter 6). This measure of complexity is model-independent; it is a measure of task complexity. As the task, in our context, is defined by the unknown distribution of the data $P(X, Y)$, information/complexity is only a measure of the compressibility of the input data, *i.e.* a measure of its underlying structural pattern or its randomness. This perspective beautifully relates to the Kolmogorov-Chaitin complexity (KC) of algorithmic information theory.

Section 8.4 revealed the overfitting term in the cross-entropy loss decomposition. The cross-entropy loss emerged *naturally* from a wishlist for representations. We shed light to Achille and Soatto insight of neutralising the overfitting term, leading to a loss function in the IB-Lagrangian form [AS18a]. This insight is the linchpin of IBT's viewpoint on generalisation.

The last *missing step* was filled in Chapter 9, where we acknowledged Chaudhari et al. demonstration that even if a deep learning model omits such regulariser term in its loss function, SGD implicitly *adds* the regulariser term [CS18; Cha+19a].

Another original consequence of the Weights-IB is that each value of the IB parameter β corresponds to a maximum (ϵ, δ) tuple of the

[AS18a] Achille and Soatto, 'Emergence of Invariance and Disentangling in Deep Representations'.

[CS18] Chaudhari and Soatto, 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'.

[Cha+19a] Chaudhari et al., 'Entropy-SGD: Biasing gradient descent into wide valleys'.

PAC tolerance and confidence margins.

10.2 ANSWERS TO RESEARCH QUESTIONS

1. **What are the fundamentals of IBT? How do they differ from the ones from MLT?** We have shown that IBT is based on IT and the IB method. IT and MLT share most assumptions (Section 5.9.2), and it is possible to bridge both subjects. MDL is an example of such bridge (Section 6.5). In terms of assumptions, the main difference is that the IB Principle (Chapter 7) assumes discrete random variables from finite spaces. Rissanen and Hinton and Van Camp have shown, however, that such a limitation is not significative, because it is possible to make the quantisation error arbitrarily small with enough resources [Ris86; HVC93].

[Ris86] Rissanen, 'Stochastic complexity and modeling'.

[HVC93] Hinton and Van Camp, 'Keeping the neural networks simple by minimizing the description length of the weights'.

2. **What is the relationship between IBT and current MLT? Are they redundant?** The IB Principle uses Shannon's theorems to define unreachable levels of tolerance-confidence, *i.e.* for a certain desired margin of tolerance ϵ, it defines the maximum confidence δ it is possible to reach and *vice-versa*. If in MLT, bias and variance are two conflicting objectives that the learning algorithm tries to minimise; IBT is a single-sided optimisation problem [Slo02] for a certain value of β (of course there is still the matter of choosing β). IBT is model-agnostic, distribution-dependent, *i.e.* generalisation is determined by the compressibility limits of the data and does not depend on the choice of a model class (in this way, it is similar to Rissanen's Stochastic Complexity [Ris86]). MLT is loss function agnostic, while the whole purpose of the IB Principle is to give a task-specific distortion measure.

[Slo02] Slonim, 'The information bottleneck: Theory and applications'.

3. **Is IBT capable of explaining the phenomena MLT already explains?** Yes, given the acceptance of an arbitrarily small quantisation error.

4. **Does IBT invalidate results in MLT?** Instead of invalidating MLT results, IBT gives new meaning to them. The pseudo-paradox evinced by Zhang et al. [Zha+16] of over parametrised models that generalise well is solved by IBT's conclusion that the complexity relates to the amount information in the parameters and not to amount of the data, the parameters themselves.

[Zha+16] Zhang et al., *Understanding deep learning requires rethinking generalization*.

5. **Is IBT capable of explaining phenomena still not well understood by MLT?** As already mentioned, IBT "rethinks" generalisation (Chapter 8) relating complexity to the data itself, instead of the model. This new paradigm provides a common narrative that allows us to give a theoretical explanation for phenomena that were only empirically understood (Section 9.5.1).

10.3 STRENGTHS, WEAKNESSES, THREATS AND OPPORTUNITIES

This section answers research questions 6 to 9.

10.3.1 *Strengths*

narrative: IBT is capable of connecting seemly unrelated phenomena (Section 9.5.1) and practices (Section 8.10) in a coherent narrative.

analysis: the usage of information measures during training "opens the black-box" of DNNs [ST17], allowing us to identify two distinct phases in training.

[ST17] Shwartz-Ziv and Tishby, 'Opening the Black Box of Deep Neural Networks via Information'.

model-independent/distribution-dependent: instead of depending on an user-defined model class, IBT depends only on the unknown data distribution, which is the task itself.

task-dependent loss: the IB Principle shows that a user-defined loss define what is relevant in the optimisation. Instead, IBT relies on the relevance variable (the target), defined by the task itself.

10.3.2 *Weaknesses*

discrete random variables in finite spaces: The IB Principle assumes discrete random variables in finite spaces. However, Rissanen and Hinton and Van Camp have already demonstrated that this is hardly a problem.

IB is ill-posed for deterministic functions: if a DNN is considered an invertible deterministic function [JSO18], the information in the activations is constant for discrete random variables (and infinite for continuous random variables). This observation seems to contradict IBT. We have shown (Chapter 6), however, that the network can be an invertible function as long as we consider the weights as our random variable and the information in the weights will bound the information in the activations

[JSO18] Jacobsen et al., 'i-RevNet: Deep Invertible Networks'.

(Chapter 8). Still, the stochastic mapping assumption during training is an overlooked consideration.

Markovian assumption: Another overlooked consideration is the Markovian assumption. IBT lacks a rigorous assessment of this assumption during training to show when it happens and why it is sufficient.

lack of rigour: IBT was presented without clear objectives: was it an analysis tool or a general theory? Also, it did not initially explain the relation between the information in the weights, for which there is a Bayesian ground, and the information in the activations, for which there is no such ground. The same lack of rigour can be seen in the overlooking of important assumptions (the Markovian assumption, for example).

10.3.3 *Threats*

discredit : IBT claims drove much attention. The lack of rigour, unfortunately, turned a natural suspicion into discredit. In Tishby's opinion, "[the critiques] are throwing the baby with the bathwater." However, the critiques were hardly unjustified. In time, a corpus of literature is corroborating with IBT's perspective and building its rigour. It is difficult to change the first impression, in any case.

fragmentation : IBT literature is still very fragmented.

10.3.4 *Opportunities*

PAC reformulation: In the PAC formulation, there is a margin of tolerance ϵ and a confidence measure δ. The IB β unifies those into a unique (ϵ, δ) limit. With that in mind, it is possible to create a PAC formulation that depends uniquely on β.

New optimisation strategies: The realisation of the fact that there are two distinct phases in training, where the macroscopic statistics abruptly change (Figure 9.2) may lead to the use of different optimisation strategies for each phase of the training.

Transfer Learning: If in IBT complexity depends uniquely on the compressibility of the input and the desired performance-generalisation level (β), it is possible to analyse the complexity of datasets and build a topology of learning tasks (as in [Ach+19]) where

[Ach+19] Achille et al., 'Task2Vec: Task Embedding for Meta-Learning'.

there is a theoretical prediction of task similarities. There is also an opportunity to relate this theoretical result to empirical findings like 'Taskonomy: Disentangling task transfer learning' [Zam+18].

[Zam+18] Zamir et al., 'Taskonomy: Disentangling task transfer learning'.

Ergodic processes: We saw that information theory does not require i.i.d. sampling (Chapter 5). We are not aware of any theoretical development in MLT that exploits this property.

Connections to Statistical mechanics: The area of Statistical Mechanics has been developed for more than a century. With the connection of machine learning and information theory, there is much to gain in exploiting findings in Statistical Mechanics in the learning realm (as did Chaudhari et al. [CS18; Cha+19a]).

[CS18] Chaudhari and Soatto, 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'.

[Cha+19a] Chaudhari et al., 'Entropy-SGD: Biasing gradient descent into wide valleys'.

10.4 CONCLUDING REMARKS

This dissertation was a "Greek endeavour" (Section 1.1.1): it tried to "connect the dots" and give ordinance to IBT Babylonian enterprise.

We found that IBT neither invalidates nor contradicts MLT, but rather conciliates MLT with Deep Learning Phenomena. IBT main weakness is its lack of rigour, a gap that is being filled with time. Interestingly, this weakness can be ascribed to a lack of assumptions definition, *i.e.* a lack of choice. The same kind of choice for which MLT is in instances criticised for (Section 4.8).

The present dissertation revealed that IBT, far from being rigorous and complete, is an emerging theory with a compelling narrative and many open opportunities for research.

Part IV

APPENDIX

A

Selected Papers in Information Bottleneck Theory

In chronological order:

1. Naftali Tishby, Fernando C. Pereira and William Bialek. 'The Information Bottleneck Method'. In: *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*. 1999, pp. 368–377

2. Ran Gilad-Bachrach, Amir Navot and Naftali Tishby. 'An Information Theoretic Tradeoff between Complexity and Accuracy'. In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf and Manfred K. Warmuth. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 595–609. ISBN: 978-3-540-45167-9

3. Ohad Shamir, Sivan Sabato and Naftali Tishby. 'Learning and generalization with the information bottleneck'. In: *Theoretical Computer Science* 411.29-30 (2010), pp. 2696–2711

4. Naftali Tishby and Noga Zaslavsky. 'Deep learning and the information bottleneck principle'. In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE. 2015, pp. 1–5

5. Alexander A Alemi, Ian Fischer, Joshua V Dillon and Kevin Murphy. 'Deep variational information bottleneck'. In: *arXiv preprint arXiv:1612.00410* (2016)

6. Ravid Shwartz-Ziv and Naftali Tishby. 'Opening the Black Box of Deep Neural Networks via Information'. In: (2017). arXiv: 1703.00810 [cs.LG]

7. Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey and David Daniel Cox. 'On the Information Bottleneck Theory of Deep Learning'. In: *International Conference on Learning Representations*. 2018

8. 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

9. Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto and Pietro Perona. 'Task2Vec: Task Embedding for Meta-Learning'. In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019

10. Alessandro Achille and Stefano Soatto. *Where is the Information in a Deep Neural Network?* 2019. arXiv: 1905.12213 [cs.LG]

11. Alessandro Achille, Matteo Rovere and Stefano Soatto. *Critical Learning Periods in Deep Neural Networks*. 2017. arXiv: 1711.08856 [cs.LG]

12. Alessandro Achille. 'Emergent Properties of Deep Neural Networks'. PhD thesis. UCLA, 2019. URL: https://escholarship.org/uc/item/8gb8x6w9

Bibliography

[Ach19] Alessandro Achille. 'Emergent Properties of Deep Neural Networks'. PhD thesis. UCLA, 2019. URL: https://escholarship.org/uc/item/8gb8x6w9 (cit. on pp. 8, 97, 126, 131, 134, 140, 168).

[Ach+19] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto and Pietro Perona. 'Task2Vec: Task Embedding for Meta-Learning'. In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019 (cit. on pp. 151, 162, 168).

[ARS17] Alessandro Achille, Matteo Rovere and Stefano Soatto. *Critical Learning Periods in Deep Neural Networks*. 2017. arXiv: 1711.08856 [cs.LG] (cit. on pp. ix, 63, 65, 147, 153, 168).

[AS18a] 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 (cit. on pp. 124, 125, 131–133, 135, 140, 144, 155, 156, 159).

[AS18b] 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 (cit. on pp. 128, 130, 131, 135–137, 140, 145, 148, 150, 168).

[AS19] Alessandro Achille and Stefano Soatto. *Where is the Information in a Deep Neural Network?* 2019. arXiv: 1905.12213 [cs.LG] (cit. on pp. 134, 140, 168).

[Aft+01] 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 (cit. on p. 69).

[AB16] Guillaume Alain and Yoshua Bengio. 'Understanding intermediate layers using linear classifier probes'. In: *arXiv preprint arXiv:1610.01644* (2016) (cit. on p. 62).

[Ale+16] Alexander A Alemi, Ian Fischer, Joshua V Dillon and Kevin Murphy. 'Deep variational information bottleneck'. In: *arXiv preprint arXiv:1612.00410* (2016) (cit. on pp. 129, 131, 141, 145, 148, 168).

[Ari00] Aristotle. *Aristotle: Nicomachean Ethics*. Cambridge Texts in the History of Philosophy. Cambridge University Press, 2000. DOI: 10.1017/CBO9780511802058 (cit. on p. 21).

[Ben12] 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 (cit. on p. 25).

[Blu07] 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 (cit. on p. 51).

[Cat08] Ariel Caticha. *Lectures on Probability, Entropy, and Statistical Physics*. 2008. arXiv: 0808.0012 [physics.data-an] (cit. on pp. 31, 32, 67, 68).

[Cha06] Gregory Chaitin. *Meta Math! The Quest for Omega*. Vintage Books, 2006. ISBN: 1400077974 (cit. on pp. 3, 16, 95, 102, 106, 146).

[CS18] P. Chaudhari and S. Soatto. 'Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks'. In: *2018 Information Theory and Applications Workshop (ITA)*. 2018, pp. 1–10. DOI: 10.1109/ITA.2018.8503224 (cit. on pp. x, 7, 144, 146, 150–152, 159, 163).

[Cha+19a] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun and Riccardo Zecchina. 'Entropy-SGD: Biasing gradient descent into wide valleys'. In: *5th International Conference on Learning Representations, ICLR 2017*. 2019 (cit. on pp. x, 146, 148, 159, 163).

[Cha+19b] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun and Riccardo Zecchina. 'Entropy-sgd: Biasing gradient descent into wide valleys'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.12 (2019) (cit. on p. 7).

[CHO19] Ivan Chelombiev, Conor Houghton and Cian O'Donnell. 'Adaptive Estimators Show Information Compression in Deep Neural Networks'. In: *International Conference on Learning Representations*. 2019 (cit. on p. 147).

[CT06] T. M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd ed. OCLC: ocm59879802. Wiley-Interscience, 2006. ISBN: 9780-471-2419-5-9 (cit. on pp. 81, 82, 85–88, 113).

[Dal] Dale. *Are Lorentz aether theory and special relativity fully equivalent?* Physics Stack Exchange. Dale (https://physics.stackexchange.com/users/204834/dale). eprint: https://physics.stackexchange.com/q/525808. URL: https://physics.stackexchange.com/q/525808 (cit. on p. 107).

[Den09] 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 (cit. on p. 14).

[DR17] Gintare Karolina Dziugaite and Daniel M. Roy. 'Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data'. In: *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. Ed. by Gal Elidan, Kristian Kersting and Alexander T. Ihler. AUAI Press, 2017. URL: http://auai.org/uai2017/proceedings/papers/173.pdf (cit. on pp. 65, 137, 138).

[Far16] Karen Farrington. *The blitzed city : the destruction of Coventry, 1940*. London: Aurum Press, 2016. ISBN: 978-1781313268 (cit. on p. 2).

[Fey94] Richard Feynman. *The Character of Physical Law*. Modern Library, 1994. ISBN: 0-679-60127-9 (cit. on p. 1).

[Gar59] Martin Gardner. *Logic machines and diagrams*. McGraw-Hill Book Company, 1959 (cit. on p. 15).

[GBNT03] Ran Gilad-Bachrach, Amir Navot and Naftali Tishby. 'An Information Theoretic Tradeoff between Complexity and Accuracy'. In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf and Manfred K. Warmuth. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 595–609. ISBN: 978-3-540-45167-9 (cit. on pp. 117, 118, 167).

[GS18] Marcelo Gleiser and Damian Sowinski. 'The Map and the Territory'. In: ed. by Shyam Wuppuluri and Francisco Antonio Doria. Springer International Publishing, 2018. ISBN: 9783319724782. DOI: 10.1007/978-3-319-72478-2 (cit. on pp. 2, 95).

[Gol+19] Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Brian Kingsbury, Igor Melnyk, Nam Nguyen and Yury Polyanskiy. *Estimating Information Flow in DNNs*. 2019. URL: https://openreview.net/forum?id=HkxOoiAcYX (cit. on p. 147).

[GBC16] 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 (cit. on pp. 23, 25, 124, 125, 152, 154).

[Gé18] Aurélien Géron. *A Short Introduction to Entropy, Cross-Entropy and KL-Divergence*. [Online; Last accessed on 2020-03-08.] 5th Feb. 2018. URL: https://youtu.be/ErfnhcEV108 (cit. on pp. 76–78, 80).

[Hau88] David Haussler. 'Quantifying inductive bias: AI learning algorithms and Valiant's learning framework'. In: *Artificial intelligence* 36.2 (1988), pp. 177–221 (cit. on pp. 53–55, 57–60, 99, 100).

[Hig+17] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed and Alexander Lerchner. 'beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework'. In: *ICLR*. 2017 (cit. on p. 131).

[HVC93] Geoffrey E Hinton and Drew Van Camp. 'Keeping the neural networks simple by minimizing the description length of the weights'. In: *Proceedings of the sixth annual conference on Computational learning theory*. 1993, pp. 5–13 (cit. on pp. viii, 8, 59, 103–105, 144, 160, 161).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. 'Flat minima'. In: *Neural Computation* 9.1 (1997), pp. 1–42 (cit. on p. 153).

[HR18] 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 (cit. on p. 63).

[Hue17] Bryce Huebner. *The Philosophy of Daniel Dennett*. Oxford University Press, 2017 (cit. on p. 17).

[Hum09] David Hume. *Tratado da natureza humana*. Editora UNESP, 2009. ISBN: 97885-7139-901-3 (cit. on p. 16).

[JSO18] Jörn-Henrik Jacobsen, Arnold W. M. Smeulders and Edouard Oyallon. 'i-RevNet: Deep Invertible Networks'. In: *CoRR* abs/1802.07088 (2018). arXiv: 1802.07088 (cit. on p. 161).

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003. ISBN: 0-521-59271-2 (cit. on pp. 17, 31–33, 35).

[KT08] 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 (cit. on p. 62).

[KW14] Diederik P. Kingma and Max Welling. 'Auto-Encoding Variational Bayes'. In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014. URL: http://arxiv.org/abs/1312.6114 (cit. on pp. 136, 137).

[KSW15] Durk P Kingma, Tim Salimans and Max Welling. 'Variational Dropout and the Local Reparameterization Trick'. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett. Vol. 28. Curran Associates, Inc., 2015. URL: https://proceedings.neurips.cc/paper/2015/file/bc7316929fe1545bf0b98d114ee3ecb8-Paper.pdf (cit. on pp. 144, 149).

[Kle18] Daniel Klein. *Mighty mouse*. [Online; Published: 2018-12-19. Accessed: 2020-01-16]. 2018. URL: https://www.technologyreview.com/s/612529/mighty-mouse/ (cit. on p. 22).

[Kle74] Martin J. Klein. 'Carnot's contribution to thermodynamics'. In: *Physics Today* 27.8 (Aug. 1974), pp. 23–28. DOI: 10.1063/1.3128802 (cit. on p. 3).

[LH07] Shane Legg and Marcus Hutter. *A Collection of Definitions of Intelligence*. 2007. arXiv: 0706.3639 [cs.AI] (cit. on p. 13).

[Les+93] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus and Shimon Schocken. 'Multilayer feedforward networks with a nonpolynomial activation function can approximate any function'. In: *Neural Networks* 6.6 (1993), pp. 861–867. ISSN: 0893-6080. DOI: https://doi.org/10.1016/S0893-6080(05)80131-5. URL: https://www.sciencedirect.com/science/article/pii/S0893608005801315 (cit. on p. 154).

[LTR17] 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 (cit. on p. 61).

[LB09] Delwin T. Lindsey and Angela M. Brown. 'World Color Survey color naming reveals universal motifs and their within-language diversity'. In: *Proceedings of the National Academy of Sciences* 106.47 (2009), pp. 19785–19790. ISSN: 0027-8424. DOI: 10.1073/pnas.0910981106. eprint: https://www.pnas.org/content/106/47/19785.full.pdf. URL: https://www.pnas.org/content/106/47/19785 (cit. on p. 138).

[LS18] Zachary C. Lipton and Jacob Steinhardt. *Troubling Trends in Machine Learning Scholarship*. 2018. arXiv: 1807.03341 [stat.ML] (cit. on p. 4).

[Ly+17] Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul Grasman and Eric-Jan Wagenmakers. *A Tutorial on Fisher Information*. 2017. arXiv: 1705.01064 [math.ST] (cit. on p. 89).

[Mac02] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. USA: Cambridge University Press, 2002. ISBN: 0521642981 (cit. on pp. 69, 72, 81, 87, 90, 102–104, 107).

[Mac68] Robert Beverley MacKenzie. *The Darwinian Theory of the Transmutation of Species Examined*. J. Nisbet, 1868, p. 318 (cit. on p. 14).

[Mac92] D. Mackay. 'The Evidence Framework Applied to Classification Networks'. In: *Neural Computation* 4 (1992), pp. 720–736 (cit. on p. 153).

[Mar16] 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/ (cit. on p. 24).

[May18] Adrienne Mayor. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton University Press, 2018. ISBN: 9780-691-18351-0 (cit. on p. 14).

[McA99] David A. McAllester. 'Some PAC-Bayesian Theorems'. In: *Machine Learning* 37.3 (1999), pp. 355–363. DOI: 10.1023/a:1007618624809 (cit. on pp. 59, 60, 137).

[McA13] David A. McAllester. 'A PAC-Bayesian Tutorial with A Dropout Bound'. In: *CoRR* abs/1307.2118 (2013). arXiv: 1307.2118 (cit. on pp. 59, 137).

[MP43] 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 (cit. on pp. 18, 23).

[McM53] Brockway McMillan. 'The Basic Theorems of Information Theory'. In: *Ann. Math. Statist.* 24.2 (June 1953), pp. 196–219. DOI: 10.1214/aoms/1177729028 (cit. on p. 81).

[Mel18] Rodrigo F. Mello. *Statistical Learning Theory*. [Online;Published: 2018-10-03. Last Accessed: 2020-04-22]. Oct. 2018. URL: https://youtu.be/KTrRap4Spd0 (cit. on p. 45).

[MP18] Rodrigo F. Mello and Moacir Antonelli Ponti. *Machine learning: a practical approach on the statistical learning theory*. Springer, 2018 (cit. on pp. 41, 45, 47, 61).

[Mit80] Tom M. Mitchell. *The Need for Biases in Learning Generalizations*. Tech. rep. Rutgers University, 1980 (cit. on pp. 48, 106).

[MRT12] Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN: 9780-262-01825-8 (cit. on pp. 25, 51, 52, 57).

[O'N16] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishing Group, 2016. ISBN: 0553418815 (cit. on p. 5).

[Pie] John R. Pierce. *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications. ISBN: 0486240614 (cit. on p. 3).

[Pop04] Karl Popper. *A Lógica da Pesquisa Científica*. Trans. by Leonidas Hegenberg and Octanny Silveira. São Paulo: Cultrix, 2004 (cit. on p. 2).

[Pri10] Jose C Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010 (cit. on p. 7).

[Rah18] Ali Rahimi. *Ali Rahimi NIPS 2017 Test-of-Time Award Presentation Speech*. https://youtu.be/x7psGHgatGM. [Online; Last accessed on 2020-08-04.] 7th Mar. 2018. URL: https://youtu.be/x7psGHgatGM (cit. on p. 1).

[Ris86] Jorma Rissanen. 'Stochastic complexity and modeling'. In: *The annals of statistics* (1986), pp. 1080–1100 (cit. on pp. viii, 102, 103, 106, 160, 161).

[RND10] 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 (cit. on pp. 4, 13, 18, 20–22, 26).

[Sax+18] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey and David Daniel Cox. 'On the Information Bottleneck Theory of Deep Learning'. In: *International Conference on Learning Representations*. 2018 (cit. on pp. 5, 146, 147, 168).

[Sax16] John G. Saxe. *The blind men and the elephant*. Enrich Spot Limited, 2016 (cit. on p. 19).

[SST10] Ohad Shamir, Sivan Sabato and Naftali Tishby. 'Learning and generalization with the information bottleneck'. In: *Theoretical Computer Science* 411.29-30 (2010), pp. 2696–2711 (cit. on pp. 97, 117, 167).

[Sha48] Claude E. Shannon. 'A mathematical theory of communication'. In: *Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on pp. 3, 69–71, 81, 88, 112).

[SW49] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949. ISBN: 978-0-252-72548-7 (cit. on pp. 71, 73, 74, 116).

[STR18] 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://youtu.be/m8PLzDmW-TY (cit. on p. 45).

[STW97] John Shawe-Taylor and Robert C Williamson. 'A PAC analysis of a Bayesian estimator'. In: *Proceedings of the tenth annual conference on Computational learning theory*. 1997, pp. 2–9 (cit. on p. 59).

[ST17] Ravid Shwartz-Ziv and Naftali Tishby. 'Opening the Black Box of Deep Neural Networks via Information'. In: (2017). arXiv: 1703.00810 [cs.LG] (cit. on pp. 5, 62, 135, 145–147, 153, 155, 161, 168).

[Slo02] Noam Slonim. 'The information bottleneck: Theory and applications'. PhD thesis. Hebrew University, 2002 (cit. on pp. 51, 111, 117, 160).

[ST19] 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 (cit. on p. 63).

[SG17] Jimmy Soni and Rob Goodman. *A mind at play: how Claude Shannon invented the information age*. Simon and Schuster, 2017 (cit. on pp. 22, 69).

[Sow16] Damian Radoslaw Sowinski. 'Complexity and stability for epistemic agents: The foundations and phenomenology of configurational Entropy'. PhD thesis. 2016 (cit. on pp. 30–32, 67, 68).

[Sto15] James V. Stone. *Information theory: a tutorial introduction*. Sebtel Press, 2015 (cit. on pp. 69, 73, 75, 76, 78, 84, 89).

[Sza11] László E. Szabó. 'Lorentzian Theories vs. Einsteinian Special Relativity — A Logico-empiricist Reconstruction'. In: *Der Wiener Kreis in Ungarn / The Vienna Circle in Hungary*. Springer Vienna, 2011, pp. 191–227. DOI: 10.1007/978-3-7091-0177-3_9 (cit. on p. 107).

[TD15] Alexander Terenin and David Draper. 'Cox's Theorem and the Jaynesian Interpretation of Probability'. In: (2015). arXiv: 1507.06597 [math.ST] (cit. on p. 32).

[Tis17a] Naftali Tishby. *Information Theory of Deep Learning*. https://youtu.be/bLqJHjXihK8. [Online; Published: 22017-08-03. Last Accessed: 2020-06-01]. 3rd Aug. 2017. URL: https://youtu.be/bLqJHjXihK8 (cit. on p. 5).

[Tis17b] Naftali Tishby. *Information Theory of Deep Learning*. https://youtu.be/FSfN2K3tnJU. [Online; Published: 2017-10-16. Last Accessed: 2020-06-01]. 16th Oct. 2017. URL: https://youtu.be/FSfN2K3tnJU (cit. on pp. 135, 149).

[Tis20] Naftali Tishby. *The Information Bottleneck View of Deep Learning: Why do we need it?* https://youtu.be/utvIaZ6wYuw. [Online; Last accessed on 2021-03-12.] 10th Jan. 2020. URL: https://youtu.be/utvIaZ6wYuw (cit. on pp. 24, 97, 145, 147, 148).

[TPB99] Naftali Tishby, Fernando C. Pereira and William Bialek. 'The Information Bottleneck Method'. In: *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*. 1999, pp. 368–377 (cit. on pp. 111, 113–115, 118, 129, 139, 167).

[TZ15a] Naftali Tishby and Noga Zaslavsky. 'Deep learning and the information bottleneck principle'. In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE. 2015, pp. 1–5 (cit. on pp. vii, 135, 148, 167).

[TZ15b] Naftali Tishby and Noga Zaslavsky. 'Deep learning and the information bottleneck principle'. In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE. 2015, pp. 1–5 (cit. on p. 5).

[Tur36] A. M. Turing. 'On Computable Numbers, with an Application to the Entscheidungsproblem'. In: *Proceedings of the London Mathematical Society* s2-42.1 (1936), pp. 230–265. DOI: 10.1112/plms/s2-42.1.230 (cit. on pp. 102, 106).

[Tur07] Alan M. Turing. 'Computing Machinery and Intelligence'. In: *Parsing the Turing Test*. Springer Netherlands, Nov. 2007, pp. 23–65. DOI: 10.1007/978-1-4020-6710-5_3 (cit. on p. 14).

[Uzg20] William Uzgalis. 'John Locke'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University, 2020. URL: https://plato.stanford.edu/archives/spr2020/entries/locke/ (cit. on p. 16).

[Val84] 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 (cit. on p. 51).

[Valoo] Harri Valpola. 'Bayesian Ensemble Learning for Nonlinear Factor Analysis'. PhD thesis. Espoo: Helsinki University of Technology, 2000 (cit. on pp. 102, 103, 106).

[VLS11] 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 (cit. on pp. 47, 58).

[WB68] C. S. Wallace and D. M. Boulton. 'An Information Measure for Classification'. In: *The Computer Journal* 11.2 (Aug. 1968), pp. 185–194. ISSN: 0010-4620. DOI: 10.1093/comjnl/11.2.185. eprint: https://academic.oup.com/comjnl/article-pdf/11/2/185/1075925/11-2-185.pdf (cit. on p. 103).

[Was13] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013 (cit. on pp. 34, 41, 42).

[Wie82] 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 (cit. on pp. 33, 64, 153).

[Wol17] Natalie Wolchover. *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/. Sept. 2017 (cit. on p. 5).

[WM97] 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 (cit. on p. 55).

[ZEASS20] Abdellatif Zaidi, Iñaki Estella-Aguerri and Shlomo Shamai (Shitz). 'On the Information Bottleneck Problems: Models, Connections, Applications and Information Theoretic Views'. In: *Entropy* 22.2 (Jan. 2020), p. 151. ISSN: 1099-4300. DOI: 10.3390/e22020151 (cit. on pp. 115, 116).

[Zam+18] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik and Silvio Savarese. 'Taskonomy: Disentangling task transfer learning'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3712–3722 (cit. on pp. x, 150, 163).

[Zas+18] Noga Zaslavsky, Charles Kemp, Terry Regier and Naftali Tishby. 'Efficient compression in color naming and its evolution'. In: *Proceedings of the National Academy of Sciences* 115.31 (2018), pp. 7937–7942 (cit. on pp. 75, 138, 139).

[Zha+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht and Oriol Vinyals. *Understanding deep learning requires rethinking generalization*. 2016. arXiv: 1611.03530 [cs.LG] (cit. on pp. v, vii, 4, 5, 27, 62, 65, 131, 140, 160).

[Zho+19] 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 (cit. on p. 62).