Université de Montréal

L'information algorithmique en physique
Émergence, sophistication et localité quantique

par

Charles Alexandre Bédard

Département de physique
Faculté des arts et des sciences

Thèse présentée à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Philosophiæ Doctor (Ph.D.)
en physique

janvier 2020

Université de Montréal

Faculté des études supérieures et postdoctorales

Cette thèse intitulée

L'information algorithmique en physique

Émergence, sophistication et localité quantique

présentée par

Charles Alexandre Bédard

a été évaluée par un jury composé des personnes suivantes :

Manu Paranjape
(président-rapporteur)

Gilles Brassard
(directeur de recherche)

Luc Vinet
(co-directeur)

Pierre McKenzie
(membre du jury)

Charles H. Bennett
(examinateur externe)

Pavel Winternitz
(représentant du doyen de la FESP)

Thèse acceptée le :
Le 26 novembre 2019

Sommaire

Cette thèse explore des aspects du monde naturel par la lentille de l'information algorithmique. La notion de l'émergence, intuitivement reliée à tant de phénomènes naturels, se voit offrir une définition cadrée dans le domaine plus spécifique des statistiques algorithmiques. Capturant toutes deux l'organisation non triviale d'un objet, la sophistication et la profondeur logique sont relativisées à un objet auxiliaire puis remises en relation. Enfin, des modèles proposant une description locale des systèmes quantiques sont démontrés équivalents, ont leur coût de description quantifié et sont généralisés aux systèmes continus.

Mots-clés

- Sophistication
- Profondeur logique
- Probabilité d'arrêt
- Émergence
- Fonction de structure
- Formalisme de Deutsch-Hayden
- Vies parallèles

Summary

This thesis explores aspects of the physical world through the lens of algorithmic information. The notion of emergence, intuitively linked to many natural phenomena, is offered a definition framed in the field of algorithmic statistics. Both capturing non-trivial organization of an object, sophistication and logical depth are compared once relativized to an auxiliary object. Finally, models proposing a local description of the quantum systems are shown equivalent, have their description cost quantified and are generalized to continuous systems.

Key Words

- Sophistication
- Logical Depth
- Halting Probability
- Emergence
- Kolmogorov's Structure Function
- Deutsch-Hayden's Formalism
- Parallel Lives

Table des matières

Sommaire v

Summary vii

Liste des figures xv

Remerciements xix

Chapitre 1. Introduction (français) 1

«The Unreasonable Effectiveness» de l'informatique... 1

Présentation de la thèse 4

Chapitre 2. Introduction (English) 7

The Unreasonable Effectiveness of Computer Science... 7

Presentation of the Thesis 10

Chapitre 3. An Algorithmic Approach to Emergence 13

3.1. Motivation 13

3.1.1. Current References to Emergence 14

3.1.2. From Systems to Bit Strings 15

3.1.3. Outline 17

3.2. A Primer on Algorithmic Methods 17

3.2.1. Algorithmic Complexity 18

Mutual Information 19

3.2.2. Nonprobabilistic Statistics ... 20
Kolmogorov's Structure Function ... 22
3.2.3. Algorithmic Connections in Physics ... 24

3.3. Defining Emergence ... 26
3.3.1. Towards a Definition ... 27
Induced Models ... 27
Induced Structure Function ... 28
Minimal Partial Models as a Signature of Emergence ... 29
3.3.2. Quantifying Emergence ... 30
The Data Specifies the Minimal Partial Models ... 30
Partial Understanding ... 31
Hierarchy of Minimal Partial Models ... 34
3.3.3. Extending Concepts ... 34
A Notion of Coarse-Graining ... 35
Boundary Conditions ... 36

3.4. Examples ... 37
3.4.1. Simulation of a 2D Gas Toy Model ... 37
3.4.2. Dynamical Systems ... 41
3.4.2.1. From Integrability to Chaos ... 41
3.4.2.2. Thermodynamics and Statistical Mechanics ... 44

3.5. Conclusion ... 46

Acknowledgements ... 49

3.6. Appendix: Proof of Theorem 10 ... 50
Proof Referring to Chapter 4 ... 50
Proof without Reference to Chapter 4 ... 50

Chapitre 4. Relativity of Depth and Sophistication ... 53

4.1. Introduction 53

4.2. Preliminaries 55

Notation 55

4.2.1. Algorithmic Complexity 55

Halting Information 57

4.2.2. Nonprobabilistic Statistics 59

The Help of $\mathbb{N}^2$ 59

Quantifying “Good” Models 60

Robustness of the Method 62

4.2.3. Logical Depth and Time-Bounded Complexity 64

4.2.4. Definitions Relativized 66

4.3. Chain Rules for Profiles 67

4.3.1. A Chain Rule for Time Profiles 67

4.3.2. Not for Description Profiles: The Antistochastic Counter-Example 69

4.4. The Gap 72

4.4.1. A Man in the Middle 72

A Computable Shape 74

4.4.2. Relativizing the $\mathcal{M}$-Profile 75

Two ways! 75

Each with his Own Mate 76

4.4.3. Losing Synchronicity 78

Relation with Halting Knowledge 79

Naming the Gap 82

4.5. Depth and Sophistication of Pairs 84

4.5.1. Depth_0 84

4.5.2. A Parameter-free Sophistication? 85

4.6. Conclusions 88

Acknowledgements 91

4.7. Appendix: Holographic Reconstruction from Time Considerations 92

Chapitre 5. Topics on Quantum Locality 93

5.1. Introduction 93

5.2. Preliminaries 95

5.2.1. Deutsch-Hayden's Formalism 95

5.2.2. Abstract Formalism of Parallel Lives 96

5.2.3. Quantum Formalism of Parallel Lives 98

5.3. Equivalences 99

5.3.1. DH's Descriptor ↔ RR's Evolution Matrix 99

5.3.2. RR: Abstract → Quantum 101

5.4. The Cost of Locality 102

5.4.1. Density-Matrix Space of a Qubit 102

5.4.2. Descriptor Space of a Qubit in an n-Qubit Network 103

5.4.3. The Universal Descriptor 104

5.4.4. What More than the Universal Wave Function? 105

5.5. Continuous Systems 107

5.6. Conclusions 108

5.7. Appendix: Introduction to the DH Formalism 110

5.7.1. A Question of Picture 110

5.7.2. Tracking Observables 111

5.7.2.1. The Descriptor of a 1-Qubit Network 112

5.7.2.2. Descriptors of an n-Qubit Network 112

5.7.2.3. DH's Shortcut 113

5.7.2.4. One more Shortcut 113

5.7.3. Evolution from the future?! 114

5.7.3.1. Back in order! 115

5.7.4. Another Evolution Equation 116

5.7.5. Not its matrix rep, but its action! 117

5.7.6. Examples 119

5.7.6.1. The CNOT 120

5.7.7. A Note to the Reader 122

5.7.7.1. Rotation on the Bloch Sphere 123

5.8. Appendix: Proofs of § 5.5 124

Chapitre 6. Conclusion (français) 127

6.1. Un bref retour sur le hasard accidentel. 128

6.2. Ω dans le monde naturel 128

La même surprise en métabiologie 129

Nos mathématiques dans le monde naturel 130

Supernova 131

Et nous? 132

Chapitre 7. Conclusion (English) 135

7.1. A Short Retake on Incidental Randomness 136

7.2. On the Naturalness of Ω 136

The Same Surprise in Metabiology 137

On the Naturalness of our *m*athematics 138

Supernova 139

And Us? 140

Bibliography 141

Liste des figures

3.1 Systems are comprehended through experimentation and observation, which yield a bit string. Models and their respective boundaries can then be defined for each string through methods from Algorithmic Information Theory. 16

3.2 Kolmogorov Structure Function 24

3.3 A structure function with many drops. 26

3.4 This figure shows schematically the essence of theorems 6, 9 and 10. Thm 6) Each minimal partial model S_i has its algorithmic information contained in the algorithmic information of x, or $K(S_i|x) = O(\log n)$. Thm 9) On the right of the picture, the bubbles represent the amount of randomness deficiency of x with respect to each model. This is shown to decrease as we go further into the minimal partial models. Thm 10) The algorithmic information of the minimal partial models is contained within each other. 30

3.5 A visual helper for the proof of Thm 9. 33

3.6 A gas particle freely moving. 38

3.7 A particle bouncing off walls. 38

3.8 Particles “collide” as if they go through one another. 38

3.9 Encoding of the configuration state into bits. 38

3.10 In black is the known upper bound of the structure function. In blue is the hypothesized real structure function 40

4.1 The anatomy of a profile $\Psi \subseteq \mathbb{N}^2$. 59

4.2 An antistochastic string understood from the description and time profiles perspectives. 70

4.3 A gap between the conditional profiles $\mathcal{L}_{y|x}$ and $\Lambda_{y|x}$ of pieces x and y of anti-stochastic strings shows that the chain rule $\mathcal{L}_{x,y} \sim \mathcal{L}_x + \mathcal{L}_{y|x}$ does not find an analogue for description profiles Λ. 71

4.4 The connection between the Ω and the Ω^z clock is given by the reach curve of z. 81

4.5 The conditional profiles and the gap between them. 81

À Ω

Remerciements

Merci Gilles! C'est grâce à la liberté et à la confiance que tu m'as accordées que j'ai pu pleinement m'épanouir en recherche. Ton émerveillement perpétuel et ton intégrité implacable sont de grandes sources d'inspiration. Merci de m'avoir confié l'enseignement de ton cours : ça a été autant un honneur qu'un plaisir.

Merci Jean-Michel pour ton écoute et ton amitié, depuis le jour 1 de cette belle aventure!
Merci Stefan pour m'avoir montrer l'importance de la créativité.
Danke Marcus and the Quitters for the refreshing time spent with all of you.
Merci Geoffroy pour ton enthousiasme scientifique contagieux.
Merci Paul pour m'avoir partagé de si belles idées.
Merci aussi à Xavier, Sophie, Louis, Luc, Yvan, Jordan, Serge-Olivier, Philippe Lamontagne, Vincent, Johannes, Ämin, Philippe Allard Guérin et tous les autres collègues et professeurs qui m'ont donné envie d'en savoir plus!

Merci papa et maman, pour votre immense soutien.
Merci Valou pour ton accompagnement fidèle dans ce projet, et bien au-delà!

Chapitre 1

Introduction
(français)

The essence of mathematics lies in its freedom.
— Georg Cantor —

Au 20^{e} siècle, trois théories majeures de la physique ont été établies : la relativité générale, la mécanique quantique et *l'informatique*. Cet ordre reflète l'utilité croissante de ces domaines pour la science et la société, et à mon avis, la profondeur de leurs enseignements sur la nature de l'univers.

«The Unreasonable Effectiveness» de l'informatique...

Paris, 1900. David Hilbert demande que les arguments utilisés dans les preuves mathématiques soient si formels, qu'un processus mécanique fini pourrait bêtement en vérifier la validité. Il espère ainsi que chaque vérité mathématique puisse être validée par une preuve dans un *système axiomatique formel* (SAF) suffisamment complet. En 1931, Kurt Gödel [46] démontre que cet espoir est vain. Pour chaque SAF non contradictoire, il existera toujours des énoncés mathématiques vrais, mais improuvables selon le SAF. C'est *l'incomplétude*.

Pour la plupart des mathématiciens l'incomplétude est un mauvais rêve, une sorte d'anomalie qui ne peut être pointée du doigt que dans des contextes extrêmement étranges et non naturels. Pas pour Alan Turing. En 1936, il exhibe [74] un nombre réel *incalculable*, soit qu'aucune procédure ne peut permettre à un mathématicien de calculer ses décimales. Pour accomplir ce tour de force, Turing *mathématise le mathématicien* dans une «machine» qui porte aujourd'hui son nom et qui a inspiré l'ordinateur moderne. Or, une telle investigation conceptuelle d'un système naturel (le mathématicien) n'est rien d'autre que de la *physique*.

Le véritable intérêt de la machine de Turing ne réside pas dans les détails précis par lesquels elle formalise le calcul. Il s'agit d'abord de mentionner l'existence de la machine universelle, qui peut simuler n'importe quelle autre machine de Turing, et, par la *thèse de Church-Turing*, n'importe quel calcul. En effet, ladite thèse stipule l'existence d'une *notion absolue de la calculabilité*, qui peut être encapsulée dans différents modèles équivalents; le modèle de Turing est l'un d'entre eux. Ainsi, n'importe quel autre engin universel d'un modèle de calcul Turing-complet peut servir de référence au calcul. Le mathématicien et l'ordinateur moderne sont, en bonne approximation, des instances physiques d'un tel engin.

Sur papier, il est facile de définir des modèles de calcul encore plus puissants que celui de Turing. Pensons par exemple à une modification de la machine de Turing qui parvient à réaliser une infinité d'étapes de calcul en une seule «grosse étape». Une telle machine pourrait résoudre le problème de l'arrêt, et donc «calculer» le fameux nombre incalculable de Turing. Mais alors, sur quelle base rejette-t-on ce genre de modèle de calcul? C'est l'apparente incapacité à construire un tel dispositif dans le monde comme on le connaît.

Ceci met en évidence *la nature physique du calcul* qui le contraint, mais aussi le dote. Par exemple, Feynman [40] signale un aspect négligé par Turing: le ruban de calcul peut être fait quantique! Cela conduit à *l'ordinateur quantique*, qui calcule plus efficacement — mais ne calcule pas plus — que la machine de Turing. Dans le même esprit, Deutsch [30] promeut la thèse de Church-Turing à un principe physique. Si les calculs et les processus physiques sont identifiés, alors *n'importe quel processus physique peut être simulé, avec précision arbitraire, par un engin universel*. Par conséquent, la «notion absolue de calculabilité» de la thèse originale ne descend pas d'un monde platonique, mais bien du nôtre.

La croyance falsifiable que chaque dynamique permise dans l'univers est en correspondence avec une dynamique possible d'un engin universel — qui peut lui-même être construit au sein de l'univers — est un principe cosmique tellement profond qu'à mon avis, l'informatique théorique est la plus grande théorie physique du siècle dernier.

La théorie algorithmique de l'information (TAI ou AIT en anglais) voit le jour dans les années 60 comme un effort conjoint [73, 51, 23] de formaliser l'induction, l'information et l'aléatoire grâce aux méthodes proposées par Turing. La théorie de l'information de Shannon [71] quantifie l'information (l'entropie) d'un objet par l'incertitude du processus probabiliste sous-jacent hypothétisé. Or, un traitement algorithmique de l'objet permet de mesurer son information en le considérant en tant que tel, exprimé par une chaîne de symboles x, sans avoir à lui imposer des «origines probabilistes». Plutôt, on lui suppose une origine calculable. Sa *complexité algorithmique* $K(x)$ est alors la longueur du plus court programme qui, donné à une machine universelle, produit ladite chaîne x. Cette mesure intrinsèque d'information acquiert un caractère universel grâce à l'habileté qu'ont les machines universelles à simuler n'importe quel processus calculable, et, par la thèse physique de Church-Turing, n'importe quel processus physique.

Comme le dit Gian-Carlo Rota [68], «Success in mathematics does not lie in solving problems but in their trivialization». En ce sens, j'estime que l'une des plus belles oeuvres de Gregory Chaitin (l'un des pères fondateurs de la théorie) a été de rendre quasi évidente l'incomplétude, ce mauvais rêve non naturel pour tant de mathématiciens. Son approche est si simple[1], qu'elle pourrait figurer dans une introduction de thèse doctorale, ou encore, être le sujet de l'exemple 1.1.1 du chapitre 1.1 d'un bon livre [59] sur la TAI.

Un nouveau mode d'explication qui réduit à de simples corollaires l'un des plus grands mystères d'une époque contient à mes yeux plus de valeur qu'une solution originale mais illisible à un problème contemporain tout aussi illisible. Elle permet des fondations plus solides et plus vastes pour que l'édifice s'érige plus haut, plus droit. La théorie algorithmique de l'information offre ce nouveau mode d'explication, grâce à son habileté à faire des maths sur les maths, soit des méta-maths [25].

L'incomplétude signifie que la complexité des vérités mathématiques est infinie [26]; il est donc intenable de les contenir dans des systèmes axiomatiques de complexité finie, comme le souhaitait Hilbert. Devant cette réalisation, Chaitin suggère une approche libre, créativement anarchique et «quasi-empirique» aux mathématiques, où la

[1] Une grande simplification — ou compréhension — vient d'Emil Post, qui a remarqué que la «procédure mécanique» de vérification de preuve demandée par Hilbert permettait de *calculablement énumérer* les preuves d'un SAF. Cela signifie qu'il existe un programme qui, étant donné une description d'un SAF (par exemple, les axiomes et les règles de logique), énumère toutes les preuves du système sans jamais ne s'arrêter.

Théorème d'incomplétude de Chaitin

Pour tout système axiomatique formel $\mathcal{A}_0$ il existe une constante k_0, telle que pour presque toute chaîne de bit x, l'énoncé mathématique :

«La chaîne x est telle que $K(x) > k_0$»

est vrai, mais improuvable dans $\mathcal{A}_0$.

Preuve: D'abord, il est vrai que pour presque toute chaîne x, $K(x) \geq k_0$. En effet, il n'existe qu'un nombre fini de programmes plus courts que k_0, et chacun d'entre eux ne peut être le plus court programme que d'au plus une seule chaîne.

Ensuite, supposons par l'absurde qu'il existe une chaîne dont la complexité est *prouvablement* plus grande que k_0. Considérons alors le programme suivant.

```
$p$ :  Calculablement énumérer les preuves de $\mathcal{A}_0$
       jusqu'à trouver une preuve de
       «La chaîne $y$ est telle que $K(y) > k_0$», pour un certain $y$.
       Retourner ce $y$.
```

Le programme p peut avoir longueur $K(\mathcal{A}_0, k_0) + c \equiv k_0$ et calcule une chaîne y dont *le plus court* programme a pourtant longueur $> k_0$. Contradiction. □

complexité des systèmes formels est augmentée pour inclure des axiomes pragmatiquement justifiés par leur utilité, plutôt que par la traditionnelle auto-évidence qui semble épuisée. "To put it bluntly, from the point of view of AIT, mathematics and physics are not that different".

Présentation de la thèse

Cette thèse investigue certains aspects du monde naturel à travers la lentille de l'information algorithmique. Elle se déploie en trois articles.

An Algorithmic Approach to Emergence

CA Bédard et Geoffroy Bergeron

Cet article propose une définition quantitative de l'émergence, intuitivement reliée à tant de phénomènes naturels. Notre proposition utilise la théorie algorithmique de l'information — plus précisément les statistiques non probabilistes — pour faire un constat objectif de la notion. L'émergence seraient marquée par des sauts de la fonction de structure de Kolmogorov. Cette définition offre des résultats théoriques en plus d'une extension des notions de «coarse-graining» et de conditions frontières. Nous confrontons finalement notre définition à des applications aux systèmes dynamiques et à la thermodynamique.

L'article est écrit pour des lecteurs provenant de plusieurs disciplines. En ce sens, aucune connaissance technique en théorie algorithmique de l'information n'est préalable. Bien que les détails soient fournis, il n'est pas essentiel de tous les comprendre entièrement pour tout de même suivre le fil des idées. Nous envisageons de le publier dans un journal de physique.

Relativity of Depth and Sophistication

CA Bédard

La profondeur logique et la sophistication sont deux mesures quantitatives de l'organisation non triviale d'un objet. Bien qu'apparemment différentes, ces mesures ont été prouvées équivalentes, lorsque la profondeur logique est renormalisée par busy beaver. Dans cet article, les mesures sont relativisées à de l'information auxiliaire et il est démontré que l'habileté à résoudre le problème d'arrêt à partir de l'information auxiliaire introduit un déphasage entre les mesures. Finalement, similairement à la complexité algorithmique, la sophistication et la profondeur logique (renormalisée) offrent chacune une relation entre leur expression de (x,y), (x) et $(y|x)$.

Cet article est plus technique et demande une certaine maîtrise de divers concepts en théorie de la calculabilité et de la théorie algorithmique de l'information. Je compte le publier dans un journal d'informatique.

Topics on Quantum Locality

CA Bédard

Il y maintenant 20 ans que Deutsch et Hayden ont démontré l'existence d'une description locale et complète des systèmes quantiques. Plus récemment, Raymond-Robichaud a proposé une autre approche. Ces modes de description des états quantiques sont d'abord montrés équivalents. Ils ont ensuite leur coût de description quantifié par la dimensionnalité de leur espace : la dimension d'un seul qubit croît exponentiellement avec la taille du système total considéré. Finalement, les méthodes sont généralisées aux systèmes continus.

Cet article touche à des concepts de la théorie quantique tant par une approche physicienne qu'informaticienne. J'espère toutefois qu'il ne soit pas nécessaire d'avoir ces deux bagages pour bien le comprendre, c'est-à-dire, que tant les informaticiens que les physiciens pourront y trouver leur compte. Je fais un effort pédagogique en incluant une annexe qui réexplique le formalisme de Deutsch et Hayden avec davantage d’accompagnement.

Je vous souhaite bonne lecture.

Chapitre 2

Introduction

(English)

The essence of mathematics lies in its freedom.

— Georg Cantor —

In the 20th century, three major theories of physics have been established: general relativity, quantum mechanics and *computer science*. This order reflects the increasing usefulness of these areas for science and society, and in my opinion, the depth of their teachings on the nature of the Universe.

The Unreasonable Effectiveness of Computer Science...

Paris, 1900. David Hilbert asks that the arguments used in mathematical proofs be so formal, that a finite mechanical process could brainlessly verify their validity. He thus hopes that every mathematical truth can be validated by a proof in a sufficiently complete *formal axiomatic system* (FAS). In 1931, Kurt Gödel [46] demonstrates that this hope is futile. For each logically consistent FAS, there will always be mathematical statements that are true, but unprovable in the FAS. This is *incompleteness*.

For most mathematicians, incompleteness is a bad dream, some kind of anomaly that can only be pointed out in utterly unnatural contexts. Not for Alan Turing. In 1936, he exhibits [74] an *uncomputable* real number, *i.e.*, no procedure exists for a mathematician to compute its decimals. To accomplish this tour de force, Turing *mathematizes the mathematician* in a "machine" that today bears his name and inspired the modern computer.

Such a conceptual investigation of a natural system (the mathematician) is nothing else than *physics*.

The real interest of Turing machines does not reside in the precise details by which it formalizes computation; Rather, it lies in the existence of a universal Turing machine, which can simulate any other Turing machine and, by the *Church-Turing thesis*, any computation. Indeed, this thesis stipulates the existence of an *absolute notion of computability*, which can be encapsulated in different equivalent models; Turing's model is one of them. Hence, any other universal device of a Turing-complete computation model can serve as a computational reference. The mathematician and the modern computer are, in good approximation, physical instantiations of such devices.

On paper, it is easy to define computation models even more powerful than that of Turing. For example, consider a modification of the Turing machine which manages to carry out an infinite number of computation steps in a single "big step". Such a machine could solve the halting problem, and thus "compute" Turing's famous uncomputable number. But then, what is the basis for rejecting this kind of computation model? Simply put, it is the apparent inability to actually build a device as such in the world as we know it.

This highlights *the physical nature of computation*, which contrains it but also gifts it. For instance, Feynman [40] points out an aspect neglected by Turing: the computation tape can be made quantum! This leads to *the quantum computer*, which computes more efficiently — but does not compute more — than the Turing machine. In the same spirit, Deutsch [30] deepens the Church-Turing thesis to a physical principle. If computations and physical processes are identified, then *any physical process can be simulated, with arbitrary precision, by a universal device*. Hence, the "absolute notion of computability" of the original thesis does not come from a platonic world, but from ours.

The falsifiable belief that *any* possible dynamics, happening anywhere in the Universe, is in one to one correspondence with the possible dynamics of some fixed universal device — a device that can be built *within* the Universe — is such a deep cosmic principle that in my opinion, theoretical computer science is the most profound physical theory of the last century.

Algorithmic information theory (AIT) was born in the 1960s as a joint effort [73, 51, 23] to formalize induction, information and randomness thanks to methods proposed by Turing. Shannon's information theory [71] quantifies the information (entropy) of an object by the uncertainty of the hypothesized underlying probabilistic process. However, an algorithmic consideration of the object enables to measure its information by considering it as such, expressed by a string of symbols x, without having to impose on it any "probabilistic origins". Rather, it is assumed to have a computable origine. Its *algorithmic complexity* $K(x)$ is then the length of the shortest program which, given to a universal device, produces the string x. This intrinsic measure of information acquires a universal character from the ability of universal machines to simulate any computable process, and, by the physical Church-Turing principle, any physical process.

Gian-Carlo Rota [68] has been quoted "Success in mathematics does not lie in solving problems but in their trivialization". In this sense, I think that one of the most beautiful work of Gregory Chaitin, a founding father of the theory, was to make incompleteness — this unnatural bad dream — look almost self-evident. His approach is so simple[1], that it could be presented in the introduction of a doctoral thesis, or else be the subject of example 1.1.1 of chapter 1.1 of a good book [59] on AIT.

A new mode of explanation which reduces to mere corollaries some of the greatest mysteries of the past contains to me more value than an original but illegible solution to an equally illegible problem of today. It allows for stronger and broader foundations, so the building can rise higher, more upright. AIT offers this new mode of explanation by its ability to do mathematics on mathematics, or meta-maths [25].

Through the algorithmic lens, incompleteness means that the complexity of mathematical truths is infinite [26]; It is therefore untenable to imprison them in an axiomatic system of finite complexity, as Hilbert wished. Chaitin then suggests a free, creatively anarchic and "quasi-empirical" approach to mathematics. One in which the complexity of formal systems increases to include new axioms pragmatically justified by their utility,

[1] A big simplification — or understanding — comes from Emil Post, who noticed that Hilbert's "mechanical procedure" of proof verification entails to "computationally enumerate" the proofs of any FAS. This means that there is a program that, given a description of a FAS (*e.g.*, the axioms and the rules of logic), lists all the proofs of the system without ever stopping.

rather than the traditional self-evidentness that seems exhausted. “To put it bluntly, from the point of view of AIT, mathematics and physics are not that different.”

Chaitin’s Incompleteness Theorem

For any formal axiomatic system $\mathcal{A}_0$ there is a constant k_0 such that for almost all bit string x, the mathematical claim:

“The string x is such that $K(x) > k_0$”

is true, but unprovable in $\mathcal{A}_0$.

Proof: First, it is true that for almost all string x, $K(x) \geq k_0$. Indeed, there is only a finite number of programs shorter than k_0, and each of them can be the shortest program of no more than one string.

Then, suppose *ad absurdum* that there is a string whose complexity is *provably* greater than k_0. Then consider the following program.

```
$p$:  Computably enumerate all the proofs of $\mathcal{A}_0$
      until is found a proof of
      “The string $y$ is such that $K(y) > k_0$”, for some $y$.
      Return that $y$.
```

The program p can be made of length $K(\mathcal{A}_0, k_0) + c \equiv k_0$ and computes a string y whose *shortest* program yet has length $> k_0$. Contradiction. □

Presentation of the Thesis

This thesis investigates some aspects of the natural world through the lens of algorithmic information. It unfolds in three articles.

An Algorithmic Approach to Emergence

CA Bédard and Geoffroy Bergeron

This article proposes a quantitative definition of *emergence*. Our proposal uses algorithmic information theory — more precisely nonprobabilistic statistics — to make an objective statement of the notion. Emergence would be marked by jumps of the Kolmogorov structure function. Our definition offers some theoretical results, in addition to an extension of the notions of coarse-graining and boundary conditions. Finally, we confront our definition with applications to dynamical systems and thermodynamics.

The article is written for readers from several disciplines. In this sense, no technical knowledge in algorithmic information theory is required. Although details are provided, they are not essential to follow the thread of ideas. We plan to publish it in a physics journal.

Relativity of Depth and Sophistication

CA Bédard

Logical depth and sophistication are two quantitative measures of the non-trivial organization of an object. Although apparently different, these measures have been proven equivalent, when the logical depth is renormalized by the busy beaver function. In this article, the measures are relativized to auxiliary information and re-compared to one another. The ability of auxiliary information to solve the halting problem introduces a distortion between the measures. Finally, similar to algorithmic complexity, sophistication and logical depth (renormalized) each offer a relation between their expression of (x,y), (x) and $(y|x)$.

This article is more technical and requires some ease with various concepts of computability theory and algorithmic information theory. I intend to publish it in a computer science journal.

Topics on Quantum Locality

CA Bédard

It has been 20 years since Deutsch and Hayden demonstrated that quantum systems can be completely described locally — notwithstanding Bell theorem. More recently, Raymond-Robichaud proposed another approach to the same conclusion. First, these means of describing quantum systems are shown to be equivalent. Then, they have their cost of description quantified by the dimensionality of their space: The dimension of a single qubit grows exponentially with the size of the total system considered. Finally, the methods are generalized to continuous systems.

This article deals with concepts of quantum theory through both a physicist and a computer scientist standpoint. However, I hope that it is not necessary to have the two backgrounds to understand it well, that is to say, that both computer scientists and physicists can find their account. I make a pedagogical effort by including an appendix that reexplains the formalism of Deutsch and Hayden with more details.

Thank you for reading my dissertation. I wish you a good journey.

Chapitre 3

An Algorithmic Approach to Emergence

Abstract. This article proposes a quantitative definition of *emergence*. Our proposal uses algorithmic information theory — more precisely nonprobabilistic statistics — to make an objective statement of the notion. Emergence would be marked by jumps of the Kolmogorov structure function. Our definition offers some theoretical results, in addition to an extension of the notions of coarse-graining and boundary conditions. Finally, we confront our definition with applications to dynamical systems and thermodynamics.

3.1. Motivation

Emergence is a concept often referred to in the study of complex systems. Coined in 1875 by the philosopher George H. Lewes in his book *Problems of Life and Mind* [58], the term has ever since mainly been used in qualitative discussions [64, 12]. In most contexts, emergence refers to the phenomenon by which properties of a complex system, composed of a large quantity of simpler subsystems, are not exhibited by those simple systems by themselves, but only through their collective interactions. The following citation from Wikipedia [1] reflects this popular idea: “For instance, the phenomenon of life as studied in biology is an emergent property of chemistry, and psychological phenomena emerge from the neurobiological phenomena of living things”.

For claims such as the above to have any meaning, an agreed upon definition of emergence must be provided. Current definitions are framed around a *qualitative* evaluation of the “novelty” of properties exhibited by a system with respect to those of its constituent subsystems. This state of matters renders generic use of the term ambiguous and subjective, hence problematic within a scientific discussion. In this paper, we attempt to free the notion of emergence from subjectivity by proposing a mathematical, hence *quantitative*, notion of emergence.

3.1.1. Current References to Emergence

We review a few of the many appeals to the notion of emergence. One of them goes all the way back to Aristotle's metaphysics [67]:

> *The whole is something over and above its parts, and not just the sum of them all...*

This famous — almost pop culture — idea is revisited by the theoretical physicist Philip W. Anderson [2], who claims that "[...] the whole becomes not only more, but very different from the sum of its parts". In the same essay, he highlights the asymmetry between reducing and constructing:

> *The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. In fact, the more elementary particle physicists tell us about the nature of the fundamental laws, the less relevance they seem to have for the very real problems of the rest of science, much less to those of society.*
>
> *The constructionist hypothesis breaks down when confronted with the twin difficulty of scale and complexity. [...] at each level of complexity, entirely new properties appear, and the understanding of the new behaviours requires research which I think is as fundamental in its nature as any other. [...] At each stage, entirely new laws, concepts, and generalizations are necessary, requiring inspiration and creativity to just as great a degree as the previous one. Psychology is not applied biology, nor biology is applied chemistry.*

More recently, David Wallace [79, Chapter 2] qualifies emergent entities to be "not directly definable in the language of microphysics (try defining a haircut within the Standard Model) but that does not mean that they are somehow independent of that underlying microphysics". The notion of structures, or patterns, often related to the concept of emergence are specified by Dennett's Criterion [29] (the criterion was named by Wallace in [78]).

> Dennett's Criterion. *An emergent object exhibits patterns. The existence of patterns as real things depends on the usefulness — in particular the explanatory power and the predictive reliability — of theories which admit those patterns in their ontology.*

Dennett's criterion, when applied to the notion of temperature, tells us that it should be thought to be an emergent but real concept because it is a *useful pattern*. As Wallace [79, Chapter 2] observes, even if temperature is not a fundamental entity of the microphysics, a scientific discussion of some gas with no reference to temperature completely misses one of the most important point. In this spirit, temperature is as real as it is useful. This notion of useful patterns, or structures, is the key concept that we shall formalize in our approach.

3.1.2. From Systems to Bit Strings

To better justify the mathematical framework, we present our philosophical standpoints. We take the realist view that there is a world outside of our perception. This world is made of physical systems, and the mission of science is to understand their properties, their dynamics and their possibilities. This is done through an interplay between formulation of theories and experimental validation or falsification. Theories have the purpose of providing simple models to explain the data, a concept which will be explored throughout the paper.

Experimentation and observation, on the other hand, collect data from physical systems. The main philosophical question that we want to address here is how to get from a system, which is a real thing out there, to a string of symbols that we shall take binary?

Observation starts by an interaction between the physical system we care to learn about and some measurement apparatus. The measurement apparatus then interacts with a computing device (this can be a human) that arranges its memory in a physical representation of a bit string x. At this stage, we shall talk about the data x as its mathematical abstraction.

Importantly, a scientist who wants to get data about a system will be left with an x, which, clearly, is not only determined by the investigated system. The information in x shall reflect properties of other systems with which it has previously interacted, like the environment, the measurement apparatus and the scientist itself! As observed by Gell-Mann and Lloyd [44], this introduces several sources of arbitrariness into x, such as the level of detail of the description and the coding convention which maps the apparatus's configuration into bits. Also, the knowledge and cognitive biases of the scientist impacts *what* is being measured. For Gell-Mann and Lloyd, this arbitrariness was to be discarded in order to define the (algorithmic) information content of the object through that of x. We don't share this view, as we think that this arbitrariness inhibits a well-posed definition.

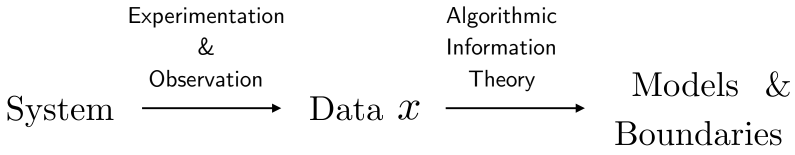


Fig. 3.1. Systems are comprehended through experimentation and observation, which yield a bit string. Models and their respective boundaries can then be defined for each string through methods from Algorithmic Information Theory.

As a side remark, we think that within this arbitrariness lies all the richness and subtleties of scientific investigation. A good scientist knows what to do with the system in order to push into x the so-far mysterious — yet to be understood — features that it can exhibit.

Nonetheless, the subjective connexion between the system and the data does not exclude an overall objective modelling of the world. For instance, if we ask a dishonest scientist to give us data about a system, but he elects instead to give us bits at whim, then investigating the data will lead to models of what was happening in the person's brain. So x is about reality, about some systems, but maybe not only about the one under investigation.

Once the data x is fixed, we face the mathematical problem of finding the best explanations for it, which is related to finding its patterns, or structures. This is the main investigation of the paper. It can be done in the realm of Algorithmic Information Theory, a branch of mathematics and logic that offers similar tools as probability theory, but with no need for unexplained randomness. Li and Vitányi, authors of the most cited textbook [59] in the field, claim that "Science may be regarded as the art of data compression". And according to the pioneer Gregory Chaitin [26], "[A] scientific theory is a computer program[1] that enables you to compute or explain your experimental data". See Figure 3.1.

One last point: Why work with classical information and computation, while the rules of the game are really quantum? The overwhelming robustness of classical information, compared to its quantum counterpart, has so far selected the former medium for information records and computations. Note that this has not prevented us from discovering

[1] Even theoretical physicists working only with pen and paper should still keep in mind that they follow some rules of symbolic manipulations that boil down to an algorithm, or a computer program.

the quantum nature of the world, and since quantum computation can be classically emulated [40], the quantum gain is only in speed, and not fundamental in terms of what we can or cannot compute. This work is grounded in computability theory, so by leaving aside questions of time complexity, we also leave aside quantum computation.

3.1.3. Outline

This paper is organized as follows. In section 2, we give a review of the basic notions of Algorithmic Information Theory, with a particular focus on nonprobabilistic statistics and connexions in physics. Building on those, we introduce in section 3 an algorithmic definition of emergence and we derive from it some concepts and results. Finally, we illustrate the relevance of the proposed definition by discussing its uses in section 4 through examples. A brief conclusion follows.

3.2. A Primer on Algorithmic Methods

Algorithmic Information Theory (AIT) originates [73, 51, 23] from the breeding between Shannon's theory of information [71] and Turing's theory of computation [74]. Titled "A Mathematical Theory of Communication", Shannon's theory concerns the ability to communicate a message that comes from a random source. This randomness, formalized in the probabilistic setting, represents ignorance, or unpredictability, of the symbols to come. The entropy is then a functional on the underlying distribution that quantifies an optimal compression of the message. Concretely, this underlying distribution is estimated through the observed biases in the frequency of the sequences of symbols to transmit. However, noticing such biases is only a single way to compress a message. For instance, if Alice were to communicate the 10^{10} first digits of π to Bob, a straightforward application of Shannon's information theory would be of no help since the frequency of the symbols to transmit is uniform (if π is normal, which it is conjectured to be). However, Alice could simply transmit:

$$\texttt{`The first } 10^{10} \texttt{ digits of } 4\sum_{n=0}^{\infty}\frac{(-1)^n}{2n+1}\texttt{.'}$$

Bob then understands the received message as an instruction which he runs on a universal computing device to obtain the desired message. Equipping information theory with universal computation enables message compression by all possible (computable) means. As we will see, the length of the best compression of a message is a natural measure of the information contained in the message.

3.2.1. Algorithmic Complexity

We give the basic definitions and properties of algorithmic complexity. See Ref. [59, Chapters 1 – 3] for details, attributions and background on computability theory.

The *algorithmic complexity* $K(x)$ of a piece of data x is the length of the shortest computable description of x. It can be understood as the minimum amount of information required to produce x by any computable process. Per contra to Shannon's notion of information, which supposes an *a priori* random process from which the data has originated, algorithmic complexity is an *intrinsic* measure of information. Because all discrete data can be binary coded, we consider only finite binary strings (referred to as "strings" from now on), *i.e.*,

$$x \in \{0,1\}^* = \{\epsilon, 0, 1, 00, \ldots\},$$

where ϵ stands for the empty word. For a meaningful definition, we have to select a universal[2] computing device $\mathcal{U}$ on which we execute the computation to obtain x from the description, which we shall call the program p. Since p is itself also a string, its length is well defined, and noted $|p|$. Therefore,

$$K_{\mathcal{U}}(x) \overset{\text{df}}{=} \min_p \{|p| \colon \mathcal{U}(p) = x\}.$$

Note that $\mathcal{U}$ can be understood as any Turing-complete model of computation, such as Turing machines, recursive functions, or concretely, it could be a modern computer or a human with pen and paper. This is the essence of the Church-Turing's thesis, according to which, all sufficiently generic approaches to symbolic manipulations are equivalent. The reader who is unfamiliar with computability theory could think of $\mathcal{U}$ as his favourite programming language on a modern computer. The invariance theorem for algorithmic complexity guarantees that no other formal mechanism can yield an essentially shorter description. This is because the reference universal computing device $\mathcal{U}$ can simulate any

[2] In the realm of Turing machines, a universal device expects an input p encoding a pair $p = \langle q, i \rangle$ and simulates the machine of program q on input i.

other computing device $\mathcal{V}$ with a constant overhead in program length, *i.e.*, there exists a constant $C_{\mathcal{U}\mathcal{V}}$ such that

$$|K_{\mathcal{U}}(x) - K_{\mathcal{V}}(x)| \leq C_{\mathcal{U}\mathcal{V}} \tag{3.1}$$

holds uniformly for all x. In such a case, it is customary in this field to use the big-O notation[3] and write $K_{\mathcal{U}}(x) = K_{\mathcal{V}}(x) + O(1)$. Since the ambiguity in the choice of computing devices is lifted (up to an additive constant), we omit the subscript $\mathcal{U}$ in the notation. Algorithmic complexity is in this sense a *universal* measure of the complexity of x.

The *conditional algorithmic complexity* $K(x|y)$ of x relative to y is defined as the length of the shortest program to compute x, if y is provided as an auxiliary input. Then one defines

$$K(x|y) \overset{\text{df}}{=} \min_p \{|p| \colon \mathcal{U}(p,y) = x\}.$$

Multiple strings $x_1, \ldots, x_n$ can be encoded into a single one as $\langle x_1, \ldots, x_n \rangle$ The algorithmic complexity $K(x_1, \ldots x_n)$ of multiple strings is then defined as

$$K(x_1, \ldots, x_n) \overset{\text{df}}{=} \min_p \{|p| \colon \mathcal{U}(p) = \langle x_1, \ldots, x_n \rangle\}.$$

For technical reasons, we restrict the set of programs resulting in a halting computation to be such that no halting program is a prefix of another halting program, namely, the set of halting programs is a *prefix code*. One way to impose such a constraint on the programs is to have all programs to be *self-delimiting*, meaning that the computational device $\mathcal{U}$ reads the program on a separate reading tape from left to right without backing up, and halts after reading the last bit of p, but no further. This restriction is not fundamentally needed for our purposes, but it entails an overall richer and cleaner theory of algorithmic information. For instance, the upcoming relation (3.2) holds within an additive constant only if self-delimitation is imposed.

Mutual Information

One of the great achievements of Shannon's information theory is the definition of a symmetric notion of mutual information that intuitively measures how much knowing x tells us about y, or vice versa. This is also achieved in the realm of AIT.

[3] In general, $O(f(n))$ denotes a quantity that does not exceed $f(n)$ by more than a fixed multiplicative factor.

Let x^* be the[4] shortest program that computes x. Algorithmic complexity satisfies the important chain rule

$$K(x,y) = K(x) + K(y|x^*) + O(1). \tag{3.2}$$

One obvious procedure to compute the pair of strings x and y is to first compute x out of its shortest program x^*, and then use x^* to compute y, which proves the "$\leq$" part of (3.2). It turns out that this procedure is $O(1)$-close, in program length, to the optimal way of computing $\langle x, y\rangle$. This entails a symmetric definition (up to an additive constant) of *algorithmic mutual information*,

$$I(x\colon y) \overset{\text{df}}{=} K(x) - K(x|y^*). \tag{3.3}$$

3.2.2. Nonprobabilistic Statistics

Statistics as we usually know them are anchored in probability theory. Ironically, the same person who axiomatized probability theory managed to detach statistics and model selection from its probabilisitic roots. Kolmogorov suggested [52] that AIT could serve as a basis for statistics and model selection for individual data. See Ref. [76] for a modern review.

In this setting, a *model* of x is defined to be a finite set $S \subseteq \{0,1\}^*$ such that $x \in S$. It is also referred to as an *algorithmic* or *nonprobabilistic statistic*. Any model S can be quantified as small or big by its cardinality, noted $|S|$, and it can be quantified as simple or complex by its algorithmic complexity $K(S)$. To define $K(S)$ properly, let again $\mathcal{U}$ be the reference universal computing device. Let p be a program that computes an encoding $\langle x_1, \ldots, x_N\rangle$ of the lexicographical ordering of the elements of S and halts.

$$\mathcal{U}(p) = \langle x_1, \ldots, x_N\rangle, \qquad \text{where} \qquad S = \{x_1, \ldots, x_N\}.$$

Then, S^* is the shortest such program and $K(S)$ is its length. When S and S' are two models of x of the same complexity α, we say that S is *a better* model than S' if it contains fewer elements. This is because there is less ambiguity in specifying x within a model containing fewer elements. In this sense, more of the distinguishing properties of x are

[4] If there are more than one "shortest program", then x^* is the fastest, and if more than one have the same running time, then x^* is the first in lexicographic order.

Model S	Complexity	$K(S)$	Cardinality	$\log\|S\|$
$S_{\text{Babel}} = \{0,1\}^n$	Easy to describe	$K(n)+O(1)$	Large set	n
$S_x = \{x\}$	Hard to describe	$K(x)+O(1)$	Small set	0

Tab. 3.1. Complexity and cardinality of S_{Babel} and S_x

reflected by such a model. Indeed, among all models of complexity $\leq \alpha$, a model of smallest cardinality is *optimal* for this fixed threshold of complexity.

Any string x of length n exhibits two canonical models shown in Table 3.1. The first is simply $S_{\text{Babel}} = \{0,1\}^n$, which is easy to describe because a program producing it only requires the information about n. However, it is a large set, containing 2^n elements. It is intuitively a bad model since it does not capture any properties of x, except its length. The other canonical example is $S_x = \{x\}$. This time, S_x is hard to describe, namely, as hard as x is hard to describe, and it is a very tiny set with a single element. S_x also intuitively resonates as bad model, since it captures *everything* about x, even the noise or incidental randomness that significantly weighs down the description of the model. A good map of Montreal is not Montreal itself! Such modelling of noisy properties is referred to as *overfitting* in the statistics community. Kolmogorov's structure function explores trade-offs between complexity and cardinality in order to find more interesting models than S_{Babel} and S_x. We come back to it later.

If S is a model of x, then

$$K(x|S) \leq \log|S| + O(1),$$

because one way to compute x out of S is to give the $\log|S|$ bit-long index[5] of x in the lexicographical ordering of the elements of S. This trivial computation of x relative to S is known as the *data-to-model code*[6]. A string x is a *typical* element of its model S if the data-to-model code is essentially the shortest program, *i.e.*, if

$$K(x|S) = \log|S| + O(1).$$

In such a case, there are no simple properties that single out x from the other elements of S. Notice also that the data x can always be described by a *two-part description:* The

[5] In fact, the length of the index is $\lceil \log|S| \rceil$, where $\lceil\cdot\rceil$ denotes the ceiling function. Note moreover that the program can be made self-delimiting at no extra cost because the length of the index can be computed from the resource S provided.

[6] Really, it should be called model-to-data :-).

model description and the data-to-model code. Hence,

$$K(x) \leq K(S) + \log|S| + O(1)\,. \tag{3.4}$$

In his seminal paper [41] on the foundations of theoretical (probabilistic) statistics, Fischer stated: "The statistic chosen should summarize the whole of the relevant information supplied by the sample. This may be called the *Criterion of Sufficiency*.". Kolmogorov suggested an algorithmic counterpart. A model $S \ni x$ is *sufficient* for x if the two-part description with S as a model yields a nearly minimal description, *i.e.*, if

$$K(S) + \log|S| \leq K(x) + O(\log n)^{7}\,.$$

Sufficient models S play similar roles then their probabilistic versions: They are sets containing elements displaying the same structure as x. In this case, once those properties are specified, the optimal program to then single out x is to give its index in the list of all elements of S.

Finally, a good model should not give more than the relevant information supplied by the data, since we ought to favour simple models, namely, models that are easy to describe. Hence, the *minimal sufficient model* S_M is the sufficient model of minimal complexity. This model now fits best the data, without over-fitting it.

Kolmogorov's Structure Function

For a fixed string x, the trade-off between complexity and cardinality is explored by its associated structure function, which maps any complexity threshold to the log-cardinality of the optimal model S within that threshold.

Definition 1 (Structure function)**.** *The* Structure function *of a string* x, $h_x : \mathbb{N} \to \mathbb{N}$, *is defined as*

$$h_x(\alpha) = \min_{S \ni x}\{\lceil \log|S| \rceil : K(S) \leq \alpha\}.$$

We say that the (or an) α-bit optimal model *witnesses* h_x at complexity threshold α. As α increases, the witnesses will be models of decreasing cardinality, thus capturing more properties of x. Extreme points of $h_x(\alpha)$ are determined by S_{Babel} and S_x, with their

[7] Here the $O(\log n)$ refers to $K(K(S), \log|S|)$ since the self-delimited 2-part code implicitly carry the length of each part as its intrinsic information. The optimal one part code x^* in general shall not know about the size of each part.

respective complexity and cardinality presented in Table 3.1. Indeed,

$$h_x(K(n)+O(1)) \leq \log|S_{\text{Babel}}| = n \qquad \text{and} \qquad h_x(K(x)+O(1)) \leq \log|S_x| = 0\,.$$

To upper bound the function, notice one can always build a more complex model, from a previously described one, by including the first bits of index of x within the description of the model. In this case, for each bit of index specified, the log-cardinality of the resulting model reduces by one. This implies that the overall slope of the structure function must be steeper or equal to -1. Applying this argument to S_{Babel}, we conclude that the graph of $h_x(\alpha)$ is upper bounded[8] by the line $n + K(n) - \alpha$. As a lower bound, recall eq. (3.4) and apply it to the model S witnessing $h_x(\alpha)$. In such a case, $K(S) \leq \alpha$ and $\log|S| = h_x(\alpha)$, so

$$K(x) - \alpha \leq h_x(\alpha)\,.$$

This means that the graph of $h_x(\alpha)$ always sits above the line $K(x) - \alpha$, known as the *sufficiency line*. The above inequality turning into an equality (up to a logarithmic term) if and only if the witness S is a sufficient model, by definition. This sufficiency line is reached by the structure function when enough bits of model description are available to formulate a sufficient statistics for x. Once the structure function reaches the sufficiency line, it stays near it, within logarithmic precision, because it is then bounded by above and by below by the -1 slope linear regime. The sufficiency line is always reached: at the latest, S_x does it.

For concreteness, a plot of $h_x(\alpha)$, for some string x of length n, is given in Figure 3.2. In this example, the string x is such that optimal models of complexity smaller than α_M are not teaching us much about x: For each bit of model, the cardinality of the corresponding set is reduced by half, which is as (in)efficient as enumerating raw bits of x. In sharp contrast, S_M, is exploiting complex structures in x to efficiently constrain the size of the resulting set. Most likely, it will be completely different from the optimal model of $\alpha_M - 1$ bits as it will not recite trivial properties of x, but rather express some distinguishing property of the data. Indeed, from α_M bits of model, the uncertainty about x is decreased

[8] A knowledgeable reader may frown upon this simple and not-so-precise argument because prefix technicalities demand a more careful analysis as is done in [75]. Such an analysis shows that the linear relations as presented here hold up to logarithmic fluctuations.

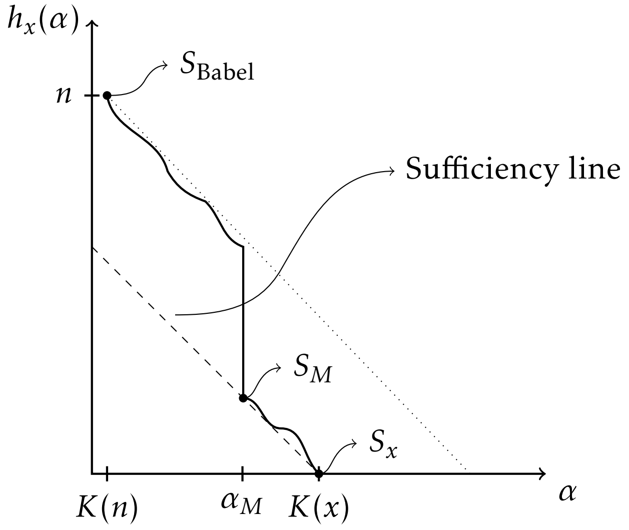


Fig. 3.2. Kolmogorov Structure Function

by much more than α_M bits, as x is now known to belong to a much smaller set. In this example, S_M is the minimal sufficient statistics.

The complexity of the minimum sufficient statistics, α_M, is known as the *sophistication* of the string x, which captures the amount of algorithmic information needed to grasp all structures — or regularities — of the string. Technically, here we refer to set-sophistication, as defined in [3], since sophistication has been originally [53] defined through total functions as model classes instead of finite sets. Importantly, Vitányi has investigated [77] three different classes of model: Finite sets, probability distributions (or statistical ensembles, *c.f.* the following section) and total functions. Although they may appear to be of increasing generality, he shows they are not. Any model of a particular class defines a model in the other two classes of the same (up to a logarithmic term) complexity and (analogue of) log-cardinality.

3.2.3. Algorithmic Connections in Physics

Ideas of using AIT and nonprobabilistic statistics to enhance the understanding of physical concepts are not new. Bennett [15] was the first to realize that thermodynamics is more a theory of computation than a theory of probability, so better rooted in AIT than in Shannon information theory. Based on his work, Zurek proposed [81] the notion of *Physical Entropy*, which generalizes thermodynamic entropy to ensue a consistent notion. An *ensemble* P is very similar to an algorithmic model for the microstate x. In general, however, a non-uniform probability distribution governs the elements of P, so the amount of information needed to specify an element $x' \in P$, on average, is given by Shannon entropy $H(P) = -\sum_{x'} P(x') \log P(x')$. Important paradoxes, such as the famous Maxwell's deamon [61, 15], rise when it is realized that the ensemble P, and hence the entropy of the system, depends upon the knowledge d held by the agent, *i.e.*, $P = P_d$. Such knowledge is usually given by macroscopic observations such as temperature, volume and pressure, and defines an ensemble P_d by the principle of maximal ignorance [66]. However, a more knowledgeable — or better equipped — agent shall gather more information d' about the microstate, which in turn defines a more precise ensemble $P' \ni x$. This leads to incompatible measures of entropy. Zurek's physical entropy S_d includes the algorithmic information contained in d as an additional tax to the overall entropy measure of the system,

$$S_d = K(d) + H(P_d).$$

Note that the similarity with Equation (3.4) is not a mere coincidence. Zurek's physical complexity encompasses a two-part description of the microstate. First, describe a model — or an ensemble — for it. Second, give the residual information to get from the ensemble to the microstate, on average. In fact, when the ensemble takes a uniform distribution over all its possible elements, Shannon's entropy $H(P)$ reduces to the log-cardinality of the ensemble, which is, up to a $k_B \ln 2$ factor, Boltzmann's entropy.

With sufficient data d, the physical entropy S_d gets close to the complexity of the microstate $K(x)$. The ensemble P_d is then analogous to a sufficient statistics, and when minimized over the complexity of d it is minimal. Baumeler and Wolf suggest [8] to take the minimal sufficient statistics (they use finite sets as in the previous section) as an objective — observer independent — statistical ensemble (they call it *the* macrostate). Gell Man and Lloyd define [44] the complexity of such a d to be the *Effective Complexity* of x. Because of Vitányi's aforementioned equivalence between finite sets and ensembles as model class, effective complexity — as defined here[9] — boils down to the same idea as sophistication, and their values are $O(\log n)$-close[10]. Müller and Szkola [6] show that

[9] Note that in Ref. [45], Gell Man and Lloyd propose to modify the definition of effective complexity. One suggestion is to keep to a fixed value some average statistical quantities that are judged to be important. Another way is by introducing a finite maximum execution time T for the universal computer to print out the bit string. But these suggestions are not deeply explored.

[10] This has been independently showed by Müller and Szkola [6, lemma 21].

strings of high effective complexity must have very large logical depth, an idea to which we shall come back.

3.3. Defining Emergence

Drops in the structure function appear to correspond to the formation of new models, which account for more properties of the observed data. In this spirit, what should one think of a string whose associated structure function is as displayed in Figure 3.3? First, does a string with such a structure function exist? In Ref. [75], it is shown that *all shapes are possible*, *i.e.*, for any graph exhibiting the necessary properties mentioned in the previous section, there exists a string whose structure function lies within logarithmic resolution of that graph.

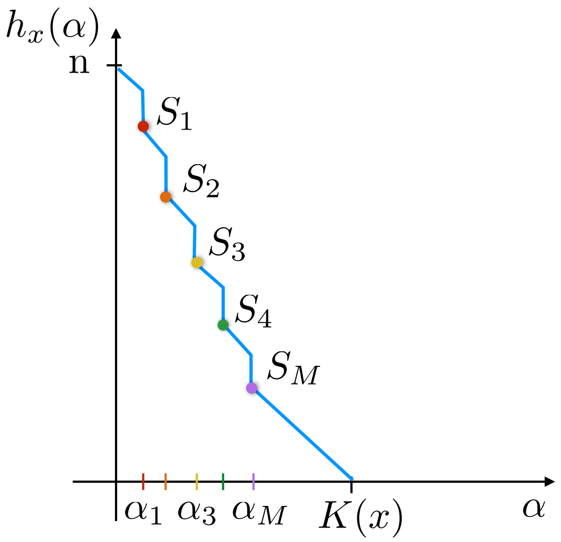


Fig. 3.3. A structure function with many drops.

It is then natural to inquire about the properties of a string with many drops in its structure function. With only a few bits of model, not much can be apprehended of x. With slightly more bits, there is a first model, S_1, capturing some useful properties of x, which leads to a more concise two-part description. And then a second model S_2, more complex, yes, but capturing even better x's properties, since the model singles out more precisely the data, yielding an overall two-part description again smaller. And so forth. Eventually, the structure function reaches the minimal sufficient statistics S_M, after which more complex models are of no help in capturing meaningful properties of x.

We now propose to relate emergence to the phenomenon by which the experimental data x exhibits a structure function with many drops. They feature structures that can be grasped at different levels of complexity and tractability.

3.3.1. Towards a Definition

In order to sharply define the models corresponding to drops of the structure function, and to make precise in which sense these are "new" and "understand" more properties, we construct a modified structure function upon which we formalize these notions.

Induced Models

As discussed briefly in the previous section, one can construct models canonically upon an already described model by including in that model, bits of the index of the data.

Definition 2 (*Induced models*)**.** *For a model $S \ni x$, the induced models $S[i]$ are given by the subset of S whose first i bits of index are the same as those of x.*

For concreteness, one way to produce such $S[i]$ is to first execute the (self-delimiting) program that computes S, and then concatenate the following (self-delimiting) program:

$$\underbrace{\texttt{The following program has } i+c\texttt{ bits.}}_{K(i)+O(1)\ \texttt{bits}}$$
$$\underbrace{\texttt{Among the strings of } S\texttt{, keep those whose index start with}}_{c\text{ bits}}\ b_1b_2b_3\ldots b_i\,. \tag{3.5}$$

where the first line of the routine is only for the sake of self-delimitation. Note that this concrete description of $S[i]$ implies

$$K(S[i]) \leq K(S) + i + K(i) + O(1).$$

Furthermore, the model $S[i] \ni x$ so defined contains half-fewer elements as S does, for every bit of index given. Hence,

$$\log|S[i]| = \log|S| - i.$$

As can be seen from Eq. (3.3.1), specifying i bits of index requires more than i extra bits of model description. Thus, we define $\delta \overset{\text{df}}{=} i + K(i) + c'$, where c' accounts for the constant size part of (3.5). A unique inverse of the relation is defined as

$$\bar{\imath}(\delta) = \max_i \{i : \; i + K(i) + c' = \delta\},$$

which represents the number of index bits that can be specified with δ extra bits of model desciption. Note that the difference between δ and $\bar{\imath}(\delta)$ is of logarithmic magnitude. We note the induced model $S(\delta) \overset{\text{df}}{=} S[\bar{\imath}(\delta)]$. Hence,

$$h_x(K(S) + \delta) \leq \log|S(\delta)| = \log|S| - \bar{\imath}(\delta).$$

Induced Structure Function

We slightly modify the structure function for reasons that will become clear later. First, we define, for each α, a model of complexity less or equal to α, which has log cardinality very close to $h_x(\alpha)$. We do so by mapping α to the[11] witness of $h_x(\alpha)$ whenever α corresponds to a drop of the structure function. And whenever the structure function is in a -1 slope regime, we map α to an induced model that builds upon the last witness of $h_x(\alpha)$. Formally, let k_0 be the smallest complexity threshold for which $h_x(k_0)$ is defined. Define the sequence of models $\{S^{(\alpha)}\}$ recursively through

$$S^{(k_0)} = S \text{ with } S \text{ a witness of } h_x(k_0) \tag{3.6}$$

$$(S^{(\alpha)}, k_\alpha) = \begin{cases} (S^{(k_{\alpha-1})}(\alpha - k_{\alpha-1}), k_{\alpha-1}) & \text{if } \log|S^{(k_{\alpha-1})}(\alpha - k_{\alpha-1})| - h_x(\alpha) = \varepsilon \\ (S, \alpha) \text{ with } S \text{ a witness of } h_x(\alpha) & \text{otherwise,} \end{cases}$$

where $\varepsilon = O(\log n)$, and can be more precisely determined in the proof of Theorem 10.

Note that the set of numbers $\{k_\alpha\}$ correspond to the set of values of α for which there are significant drops in the structure function. As can be seen from the above definition, a "significant drop" corresponds to a decrease of ε in the structure function, which is beyond what is naturally entailed by inducing the model one bit further.

Definition 3 (*Induced structure function*)**.** *The induced structure function $\tilde{h}_x(\alpha)$ is defined as*

$$\tilde{h}_x(\alpha) = \log|S^{(\alpha)}|.$$

[11] If the witness of $h_x(\alpha)$ is not unique, we choose the fastest one produced by α-bit programs.

It follows from this definition that $\tilde{h}_x$ lies just above h_x, within an additive logarithmic term smaller than ε. Why define an induced structure function $\tilde{h}_x$, which is very close to the original structure function h_x? An important difference is that the construction of the induced structure function $\tilde{h}_x(\alpha)$ in (3.6) keeps track of the actual models used at each complexity threshold. This has two advantages. First, it now becomes clear what a "drop" of the structure function is: it corresponds to a point in the construction of the induced structure function where the model used is updated rather than induced. Second, for two neighbouring points α and β in a slope -1 regime, nothing guarantees that the model witnessing $h_x(\alpha)$ and $h_x(\beta)$ are not completely different. They could *a priori* be completely different models, capturing completely different properties about the string x, but it just happens that the difference of their log-cardinality is roughly $\beta - \alpha$. On the other hand, the defining models of $\tilde{h}_x$ are constructed in a way that the -1 slope forces the models to be induced from the same original model. They simply contain more or less of the index of x. Finally, departure from the slope -1 regime in the function $\tilde{h}_x$ indicates that a new model is used, one that intuitively captures other properties of x.

Minimal Partial Models as a Signature of Emergence

We have emphasized in the construction of the induced structure function a difference between the slope -1 regime and the sharp drops of the structure function. Indeed, while the former amounts to induced models, the latter corresponds to relevant yet partial models. These will be central to our proposed definition of emergence.

Definition 4 (*Minimal partial models*)**.** *The minimal partial models are defined as the witnesses of the drops of the $\tilde{h}_x$, namely the models $\{S^{(k_\alpha)}\}_{\alpha \in 1,\dots,K(x)}$ as defined in* (3.6).

In what follows, we denote by S_1, S_2, $\dots$, S_M the successive minimal partial models with respective complexity $\alpha_1 < \alpha_2 < \dots < \alpha_M$. Minimal partial models are in some sense the interesting models, out of all the optimal models witnessing the structure function.

Definition 5 (*Emergence*)**.** *Emergence is the phenomenon characterized by observation strings that display several minimal partial models.*

Moreover, it is a function of the observation string x, not of the real object that x is supposed to represent. For instance, in the case of the dishonest scientist who disregards the object under investigation to give bits at whim, if x displays emergence, this simply

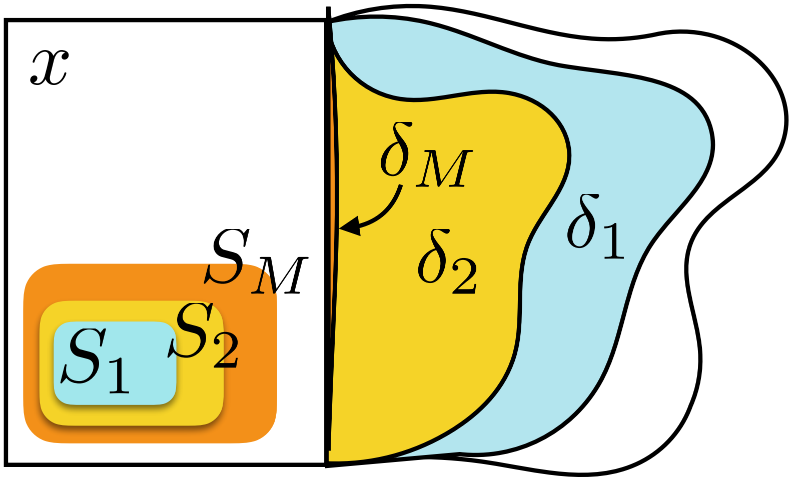


Fig. 3.4. This figure shows schematically the essence of theorems 6, 9 and 10.
Thm 6) Each minimal partial model S_i has its algorithmic information contained in the algorithmic information of x, or $K(S_i|x) = O(\log n)$. Thm 9) On the right of the picture, the bubbles represent the amount of randomness deficiency of x with respect to each model. This is shown to decrease as we go further into the minimal partial models. Thm 10) The algorithmic information of the minimal partial models is contained within each other.

means that the real objects that inspired x (surroundings, past memories, current brain state...) display emergence. Inevitable and plentiful interactions between systems make it impossible to have data about a precise object and nothing else. But that's fine, because emergence is likely the fruit of the richness of these interactions.

3.3.2. Quantifying Emergence

Under the proposed definition of emergence, we develop quantitative statements. In this section, three theorems are presented. We do not claim absolute originality, as these results use ideas that have been previously developed in algorithmic statistics [**43**, **75**, **76**]. The novelty lies in the formulation in terms of minimal partial models.

The Data Specifies the Minimal Partial Models

The first theorem confirms a basic intuition. The minimal partial models should be thought as an optimized ways to give the structural information about x, so in particular, we expect that most of their algorithmic information is in fact information *about* x. This the case.

Theorem 6. *The minimal partial models S_i can be computed from x and a logarithmic advice,*

$$K(S_i\,|\,x^*) = K(K(S_i), \lceil \log|S_i| \rceil \,|\, x^*) + O(1) = O(\log n).$$

Using the chain rule Eq. (3.2) twice allows to expand $K(x, S_i)$ in two ways:

$$\begin{aligned} K(x, S_i) &= K(x) + K(S_i\,|\,x^*) + O(1) \\ &= K(S_i) + K(x\,|\,S_i^*) + O(1), \end{aligned}$$

so another way to phrase the theorem is

$$K(x) = K(S_i) + K(x\,|\,S_i^*) + O(\log n),$$

so producing S_i in order to get x is not a waste, in fact, it is almost completely a part of x's information. See Figure 3.4.

Proof. We give a program q of length $K(K(S_i), \lceil \log|S_i| \rceil \,|\, x^*) + O(1)$ that computes S_i out of x^*.

```
q : Compute $K(S_i)$ and $\lceil \log|S_i| \rceil$ from $x^*$ (if useful)
    Run all $p$ of length $K(S_i)$ in parallel
    If $p$ halts with $\mathcal{U}(p) = \langle S \rangle$:
        If $\log|S| \leq \lceil \log|S_i| \rceil$ and $\mathcal{U}(x^*) = x \in S$:
            Print $S$ and halt.
```

□

Partial Understanding

We now justify the use of the term "partial" to qualify the nonsufficient minimal partial models. Intuitively, sharp drops of the structure function should be in correspondance with some non-trivial properties of the associated string such that the minimal model at this point should reflect an "understanding" of this properties. Naturally, the importance of this understanding could be equated to the magnitude of the drop. Theorem 9 confirms this idea, when "understanding" holds the following meaning.

Understanding, especially in AIT, amounts to reducing redundancy, as a good explanation is a simple rule that accounts for a substantial specification of the data. For instance, when one understands a grammar rule of some foreign language, she can refer to that rule to explain the many different instantiations of that rule she encounters thereafter. Those instantiations are redundant, and once the grammar rule is understood, this redundancy is reduced.

Definition 7. *The* Redundancy *of a string x of length n is defined to be*

$$Red(x) \overset{df}{=} n - K(x|n).$$

The redundancy of a string is thus the number of bits of a string that are not irreducible algorithmic information, that is, the compressible part of x. Redundancy could then be thought as a quantification of how much one understands x, once he learns x^*. Comparing x to x^*, however, is an all or nothing approach, and the whole purpose of non-probabilistic statistics is to make sense of partial understanding by studying (two-part) programs for x that interpolate between the "`print `x'`" and x^*. The next definition, in some sense, generalizes redundancy so that it can be relative to an algorithmic model.

Definition 8. *The* Randomness Deficiency *of a string x, with respect to the model S is*

$$\delta(x\,|\,S) \overset{df}{=} \log|S| - K(x|S).$$

It measures how far is x from being a typical element of the set, since a typical element would have conditional complexity as large as the size of the data-to-model code. Notice that redundancy can be recovered from

$$\delta(x\,|\,S_{\text{Babel}}) = n - K(x\,|\,S_{\text{Babel}}) = \text{Red}(x) + O(1).$$

We can then explore how much each minimal partial model reduces the randomness deficiency — or understands — the data x. Define d_i as the height of the drop just before getting to S_i, namely,

$$d_i \overset{\text{df}}{=} \tilde{h}_x(\alpha_i - 1) - \tilde{h}_x(\alpha_i).$$

Theorem 9. *The height of the i-th drop measures how much more S_i reduces the randomness deficiency, compared to S_{i-1},* i.e.,

$$\delta(x \mid S_{i-1}) - \delta(x \mid S_i) = d_i + O(\log n).$$

Proof. Using the chain rule Eq. (3.2) twice, which amounts to a bayesian inversion, and Theorem 6,

$$\begin{aligned}
\delta(x \mid S_i) &= \log|S_i| - K(x \mid S_i) \\
&= h_x(\alpha_i) - K(x) - K(S_i \mid x) + K(S_i) + O(\log n) \\
&= h_x(\alpha_i) - K(x) + \alpha_i + O(\log n). \qquad (3.7)
\end{aligned}$$

With the help of Figure 3.5, and recalling that if δ extra bits of model description are given, $\bar{\imath}(\delta) = \delta + O(\log n)$ bits of index can be given, observe that

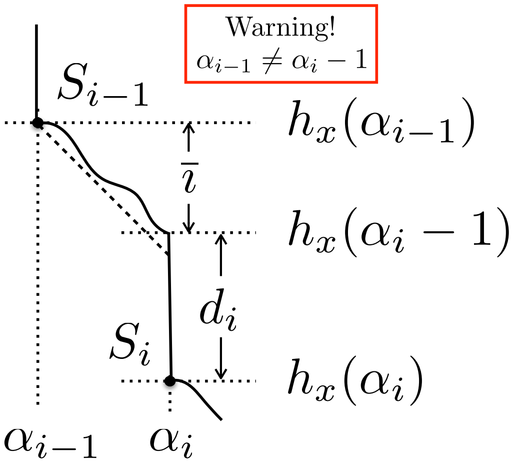


Fig. 3.5. A visual helper for the proof of Thm 9.

$$\begin{aligned}
h_x(\alpha_{i-1}) - h_x(\alpha_i) &= h_x(\alpha_{i-1}) - h_x(\alpha_i - 1) + h_x(\alpha_i - 1) - h_x(\alpha_i) \\
&= \bar{\imath}(\alpha_i - 1 - \alpha_{i-1}) + d_i \\
&= \alpha_i - \alpha_{i-1} + d_i + O(\log n).
\end{aligned}$$

Using Equation (3.7),

$$\delta(x \mid S_{i-1}) - \delta(x \mid S_i) = h_x(\alpha_{i-1}) - h_x(\alpha_i) + \alpha_{i-1} - \alpha_i + O(\log n)$$

$$= \quad d_i + O(\log n).$$

□

We can then interpret the algorithmic information in the minimal partial models as being parts of the algorithmic information of x that enable a reduction of the redundancy of x. This reduction of redundancy can be quantified by the sum of all previous drops, and the amount of redundancy left to be reduced is the sum of the drops to come. When the minimal sufficient statistic is described, α_M bits of the algorithmic information in x entail a complete reduction of the redundancy of x. The only information left to specify is then the index of $x \in S_M$, which is itself irreducible algorithmic information about x. However, this remaining information does not contain the relevant structural information about x.

Hierarchy of Minimal Partial Models

The following theorem shows that the algorithmic information in the minimal partial models is organized in a nested structure, namely, the complex minimal partial models can compute the simpler ones within logarithmic precision.

Theorem 10. *For $j > i$,*

$$K(S_i \,|\, S_j) = O(\log n).$$

Proof Sketch.

For this proof we use the results of [3], which links Set-Sophistication with Busy Beaver Logical Depth. This result implies that the shortest programs for the minimal partial models will run for so long that they are mainly consisted of Halting Information. Once it is shown that S_i and S_j are made of i and j bits (up to logarithmic error), respectively, of irreducible halting information, it becomes necessary that S_i can be computed by S_j (within logarithmic error).

The proof is relegated to Appendix 3.6.

3.3.3. Extending Concepts

We revisit the notions of coarse-graining and boundary conditions, broadening their scope.

A Notion of Coarse-Graining

Many approaches to emergence appeal to some notion of coarse-graining. For instance, when the degrees of freedom of a physical system are indexed spatially, the states correspond to functions over space. In this case, an important tool consists of averaging those state functions over regions of space, retaining only the large scale structures, as is done in the method of effective field theories. In the context of algorithmic statistics, coarse-graining is seen as a special case of what we will call *regraining*. We begin by defining the notion of coarse-graining precisely.

In set theory, coarse-grainings are defined from some mother set Ω. A *partition* $\mathcal{P}$ of Ω is a collection of disjoint and non-empty subsets such that their union gives back Ω. Let $\mathcal{P}_{\text{fine}}$ and $\mathcal{P}_{\text{coarse}}$ be two partitions of Ω. We say that $\mathcal{P}_{\text{fine}}$ is a *refinement* of $\mathcal{P}_{\text{coarse}}$ if every element in $\mathcal{P}_{\text{fine}}$ is a subset of some element of $\mathcal{P}_{\text{coarse}}$. We also say that $\mathcal{P}_{\text{coarse}}$ is a *coarse-graining* of $\mathcal{P}_{\text{fine}}$. In physics, those partitions are usually defined through non-injective functions globally defined on the state space. The pre-images of more or less informative such functions induce respectively finer or coarser partitions of the state space.

The key point of nonprobabilistic statistics is to investigate an individual object x, without needing to refer to other x' in the set of bit strings. Hence, algorithmic models are disconnected from the notion of partition, since set a single is defined specifically for x — and do not form a partition of bit strings. Still, an algorithmic model $A \ni x$ could be qualified as a *model coarse-graining* of a $B \ni x$ if $B \subseteq A$. This type of "model coarse-graining" in fact occur in the regime of induced models. However, if we compare minimal partial models to each other, even if they are of different cardinality, they are in general not subsets of one another[12]. This motivates the extended notion of *regraining*, which is simply a change from some model $A \ni x$ to $B \ni x$, where neither model is a subset of the other. It is qualified as a *fine* regraining if $|B| < |A|$ and a *coarse* one if $|B| > |A|$. Model coarse-grainings are particular cases of regrainings.

[12] Although nothing garanties that $S_j \subset S_i$, for $j > i$, S_j cannot be almost entirely composed of elements that are not in S_i. In fact, Theorem 10 states that S_i can be easily computed from S_j, so with slightly more than α_j bits of model size, an optimal model would be $S_i \cap S_j \ni x$, which, cannot be of too small cardinality unless the structure function exhibits a drop right after α_j.

The *optimal regraining* corresponds to jumping along the minimal partial models. The optimal coarse regraining occurs in the direction from S_M to S_1, and corresponds to modelling less and less precise notions of our data x to the benefit of having simpler and simpler models. It is optimal in the sense that this procedure will yield the best possible models over all possible complexities. As opposed to the traditional approaches, where the coarse-graining usually occurs in space or in time, our notion of regraining is parametrized by theory size. In some cases, properties of a physical system could be such that the optimal coarse regraining happens by averaging over space configurations, so the algorithmic regraining boils down to the traditional methods.

Boundary Conditions

Recall from §3.1.2 that an intrinsic difficulty (or beauty) of scientific investigation is that the recorded data x does not read out a single system. Even if we leave aside the effect of the measurement apparatus and the scientist on the data, it remains that systems are seldom isolated from an environment. As any interaction mediates an exchange of information, the effect of a large and complex environment will be modelled as random noise[13] in models of small complexity. But if the string x is sufficiently detailed, some structures of the environmental "noise" shall be grasped by models complex enough. This highlights that some information about x may reside outside of a simple model but inside of another; The boundary being what is relegated to the data-to-model code.

Definition 11. *The* boundary conditions of the model S *corresponding to the data* x *is the index of* x *in* S*.*

The scope of the term is broadened, so that it can be thought as the boundary of a model S, namely what, from the system that generated the observational data $x \in S$, is *not modelled* by S. This amounts to truncating the structure function after a particular minimal partial model *as if* x had no further structure to be exploited. The remaining structure in x is then viewed as coming from non-typical boundary conditions forced by interactions with an environment. In the case of the minimal sufficient statistics S_M, the

[13] An example of this situation is given by the dissipation-fluctuation theorem [21] that relates dissipative interactions in a system to the statistical fluctuations around its equilibrium point. Indeed, this theorem relates dissipation, an irreversible process that does not preserve information, with noise in the form of statistical fluctuations.

typicality of x in S_M captures the fact that the boundary conditions are arbitrary with respect to the model.

The traditional space-time boundary conditions of a system are an example of what is usually relegated to the data-to-model code, as models usually dont aim at explaining them. Another example are the precise values of mechanical friction coefficients. Within classical mechanics, these values come from outside the theory and would thus be a part of the boundary conditions. However, with more precise observations, one could explain the values of the coefficients from a more precise model that encompasses molecular interactions. More example are provided in the following section.

3.4. Examples

The versatility of the proposed approach to emergence is now illustrated through some examples. This section is not meant to be an exhaustive review of the possible uses of the proposed definitions, but should rather be understood as an illustrative complement to the main exposition.

3.4.1. Simulation of a 2D Gas Toy Model

As a first example, we consider a toy model for a 2D gas on a lattice. The gas is taken to be spatially confined on an $L \times L$ grid with a discrete time evolution. Using a pseudorandom number generator, we choose an initial position and momentum for each of the N particles. Each momentum is only a direction in the set $\{l, r, u, d\}$, corresponding to left, right, up and down. The gas then evolves according to simple rules. A single free particle, represented by a 1 in the lattice, just keeps its trajectory and momentum, as in Figure 3.6. When it bounces off a boundary, its momentum gets flipped, as in Figure 3.7, and particles collide in such a way that we can simply ignore the collision, as if they go through one another. Figure 3.8 displays examples of collisions.

At any time (including the initial time), if two or more particles are at the same site, we simply adopt a “Fock space” notation, writing down the number of particles in the site and keeping track of the momenta. As an observation x, we extract from the simulation the state in configuration space, *i.e.*, we ignore momentum, at each of the first T time steps. One visual way to encode the state in configuration space is to write in each of the

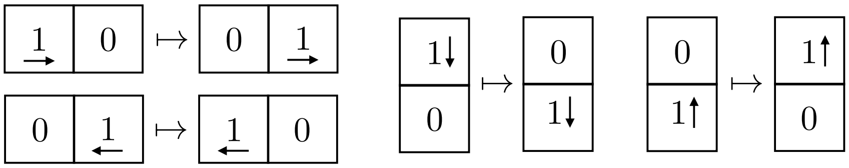


Fig. 3.6. A gas particle freely moving.

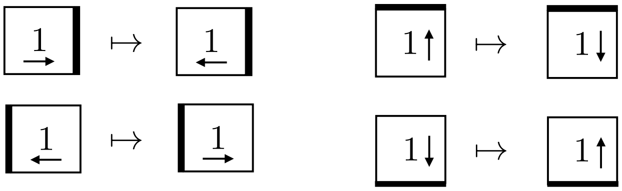


Fig. 3.7. A particle bouncing off walls.

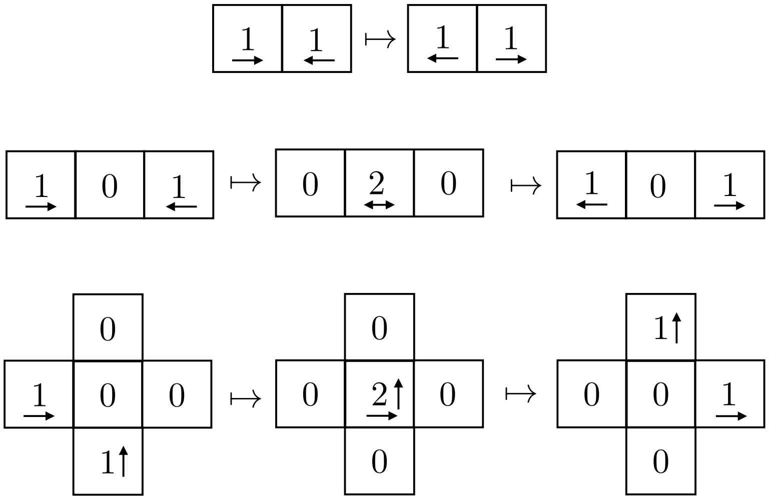


Fig. 3.8. Particles "collide" as if they go through one another.

L^2 sites a 0, and write to the left of it, in unary, the number of particles in that site. For instance, a 3×3 grid example of this coding is given in Figure 3.9.

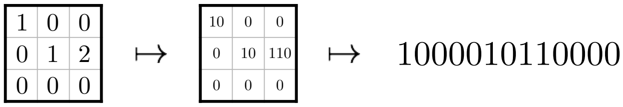


Fig. 3.9. Encoding of the configuration state into bits.

At each time step, the bit string corresponding to the configuration state has one 0 for each site of the grid, and one 1 for each particle, for a total length of $L^2 + N$. Putting all of this together, the observation x so generated is a bit string of length $|x| = (L^2 + N)T$.

Because algorithmic complexity is uncomputable, so is the structure function. However, it can be *upper semi-computed*, which means that there is an algorithm that keeps outputting better upper bounds of the structure function until it eventually reaches the actual structure function. When this happens, the algorithm does not halt, as it keeps looking for better upper bounds, not knowing that this is in vain. In our generic context of finding scientific explanations for observation data, this upper semi-computation is done by theoreticians ever finding simpler and better models. In the specific case of the simulated 2D gas, we (the theoreticians) got x from the known context of the simulation, which provides important clues to find models other than the obvious S_{Babel} and $\{x\}$.

A first model that comes from the simulation specifies the parameters L and N, external to the gas, together with the boundary time T. Compared to $\{x\}$ obtained by saving everything about the simulation, simplicity is gained by leaving open the initial conditions. This defines the set $S_{\text{Gas}}^{L,N,T}$ of all configuration histories of T iterations, for each possible initial conditions of N particles confined to a $L \times L$ grid. The size of this program is smaller than $K(L,N,T) + O(1)$, since the evolution rules are of constant length.

Even simpler models can be made by pushing into the boundaries, the particular values the external parameters L, N and T. For the illustration, we argue only with T, which we suppose to be expressed by τ bits of binary expansion. The model $S_{\text{Gas}}^{L,N,T}$ can be simplified by producing, for each possible initial condition, *all* histories of length smaller than 2^τ. We denote this set $S_{\text{Gas}}^{L,N,<2^\tau}$. Its cardinality is 2^τ times bigger, thus adding τ to the log-cardinality axis. But if T were a random number,

$$K(T) = \tau + K(\tau) + O(1), \tag{3.8}$$

exactly τ bits would be saved on the complexity axis, since only τ, as opposed to T is needed to compute the model. In general however, T is not algorithmically random, but $S_{\text{Gas}}^{L,N,<2^\tau}$ does not care about its structures, since it considers T to be outside of what is modelled. In fact, the spirit of the model is to leave completely open the boundary time parameter, so even τ could be thought as a stranger to the model. However, a threshold as such is required to define a finite set[14]. In a similar fashion, other complexity thresholds can be introduced for L and N, pushing again their complexity and/or their structures, into the boundary conditions.

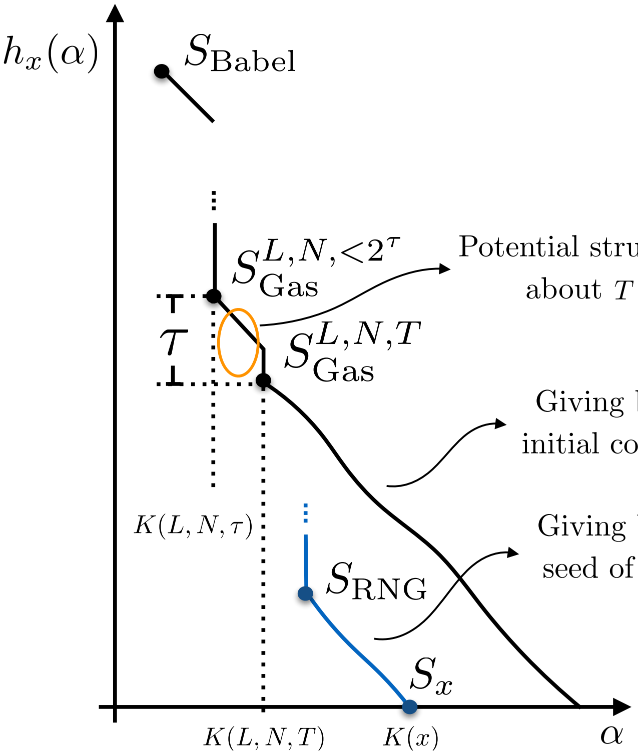


Fig. 3.10. In black is the known upper bound of the structure function. In blue is the hypothesized real structure function

As presented in Figure 3.10, the previously discussed upper bound of the structure function is likely going to be different from the real structure function. In particular, in the simulation of the gas, the initial conditions were not algorithmically random, but that they come from a pseudorandom number generator. The real structure function will grasp this fact, yielding one more drop at a later level of complexity. The witness of this drop, S_{RNG}, is then the set of all gas histories compatible with the dynamics previously described, and where the initial conditions have been generated with the pseudorandom number generator. If the slope after S_{RNG} remains in a -1 regime, it means that the seed is typical, among allowed seeds of its length. But this seed comes from another physical

[14] Taking total functions instead of finite sets as a model class eliminates this drawback. However, the possibility of taking low-complexity thresholds, like here, a power of 2, does not affect the analysis by much.

system, for instance, a human, so if the seed is long enough, the structure function could potentially find more drops that capture structures about this system.

This example makes clear that the notion of boundary conditions really refers to a theory (or an algorithmic model), and are fixed somewhat arbitrarily, when the users of the theory are satisfied with their notion of the system that is being modelled. In this case, if what we wanted to model was the gas, then $S_{\text{Gas}}^{L,N,T}$ was good enough, and it was practical to declare that the initial state was typical. But the reality may be quite different, and what we prescribe as a boundary condition to our theory may in fact be explained by a more complex, deeper theory.

3.4.2. Dynamical Systems

In this second example, we review how the notions introduced in this paper appear in the more general setting of dynamical systems. We begin by documenting how the concept of integrability and chaos can be cast in the language of algorithmic information theory. This is followed by an account of how thermodynamics can be seen to emerge, under the proposed definition of emergence, from the application of statistical mechanics to complex dynamical systems.

3.4.2.1. *From Integrability to Chaos*

Consider a generic classical system with Hamiltonian H and where the space of states M[15] is indexed by a set of real coordinates $X = \{q_i, p_i\}_i \in M$. Solutions to the dynamics are curves in M describing the evolution of the state in time. Specifying M, H and an initial point X_0 singles out a unique solution curve X_t of the dynamics. As a rudimentary formalization of some observation of the system, consider a bounded observable represented by an easy to compute function f with $f : M \to [0,1]$. From this observable, a discrete sequence is constructed from its evaluation $\{f(X_{n\tau})\}_{n\in\{1,\dots,N\}}$ at a regular time interval $0 < \tau \in \mathbb{Q}$ with $K(\tau) = O(1)$. As this sequence is to represent a series of measurements, one must restricts its resolution. Indeed, in the laboratory, as well as in numerical simulations, the values measured are always constrained to a finite resolution. For a real number α, we denote by $[\alpha]_k$ the truncation of its binary expansion after the first k bits beyond the decimal point, *i.e.*, $|[\alpha]_k - \alpha| \leq 2^{-k}$. This truncation effectively restricts the

[15] More precisely, M is a symplectic manifold parametrized locally by real coordinates forming an atlas.

resolution to k bits as the measurement function is upper-bounded by 1. Denoting by $f_n^k \equiv [f(X_{n\tau})]_k$ the restricted measurements, the recorded observational data string x is then an encoding of the sequence of measurements:

$$x \equiv \langle \{f_n^k\}_{n\in\{0,1,\ldots,N\}} \rangle,$$

where the condition that the function f be easily computed is understood as

$$K([f(X)]_k \,|\, [X]_k) = O(1).$$

We now wish to characterize the complexity of the data string x and study its asymptotic behaviour when the length N of the measurement sequence is increasing. First, one must formulate a meaningful upper bound for $K(x)$. A trivial bound is given by the bit length of the encoded sequence of measurements, thus

$$K(x) \leq kN + O(\log kN\,).$$

However, the regularity provided by physical laws implies that this bound is not strict. Indeed, given the Hamiltonian H and the manifold M, the machinery of symplectic geometry specifies the dynamical evolution as a set of differential equations that we will denote as $\langle H, M \rangle$. These equations can be integrated numerically from the initial conditions X_0 to obtain f_n to a desired precision. These remarks, together with the stated condition on f, imply[16] that

$$K(x) \leq K(\langle M, H \rangle, \tau, k, N, X_0) + O(1).$$

The above can be further simplified in view of studying the asymptotic behaviour in N by observing that the dynamical laws $\langle M, H \rangle$ and the time interval τ are fixed and independent of N. Thus, they can be taken to be constant as the length of x is scaled by increasing N such that they only contribute a constant $O(1)$ term. Hence, one has

$$K(x) \leq K(N) + K(X_0) + O(1). \tag{3.9}$$

[16] To simplify the analysis, it is tacitly assumed that the dynamical laws are simple in the sense that the coefficients of the differential equations in $\langle H, M \rangle$ are rational numbers.

Remembering that X_0 encodes the initial conditions, thus a set of real numbers which cannot be constructively specified in general, one is left with a conundrum. Indeed, if X_0 encodes typical real numbers, the upper bound (3.9) is trivial as the right-hand side becomes infinite. However, only a finite precision in the initial conditions is required in order to integrate the system to a given precision in the final result. Thus, the resolution in X_0 required is only as much as is needed to compute $\{f_n^k\}_{n\in\{0,1,\dots,N\}}$. As such, the asymptotic behaviour of $K(x)$ for $N \to \infty$ is determined by the scaling in the required resolution.

A chaotic dynamical system is often characterized by an exponential divergence of nearby initial configurations, namely

$$\frac{|X_t' - X_t|}{|X_0' - X_0|} = e^{\lambda t},$$

Where $|\cdot|$ denotes a metric on M and λ is known as the Lyapunov exponent. In such a chaotic system,

$$|X_0' - X_0| < 2^{-\ln 2\lambda n\tau - k} \qquad \Longrightarrow \qquad |X_{n\tau}' - X_{n\tau}| < 2^{-k},$$

so k bits of precision on $X_{n\tau}$ can be achieved by $k + \lambda' n$ bits of precision on X_0, where $\lambda' = \ln 2\lambda\tau$. Therefore, the computation of $X_{N\tau}$ from the initial condition is more efficient, in terms of description length, than straightforward enumeration if $k + \lambda' N \leq kN$, which is

$$\lambda' \leq k - \frac{k}{N}. \tag{3.10}$$

This means that for some values of Lyapunov exponent λ and precision k, it could more efficient to simply recite the observed data $\{f_n\}_{n\in\{0,1,\dots,N\}}$ as a genuinely random string. However, no matter how large the Lyapunov exponent is, there will always be a regime of precision for which it is more efficient to calculate $\{f_n\}_{n\in\{0,1,\dots,N\}}$ from enough bits of initial conditions. Concretely, the precision on the initial conditions that can be obtained is bounded by the resolution of measurement devices. A more practical approach accounts for this with a fixed resolution $k' > k$ in the initial conditions and is thus limited to the truncation $[X_0]_{k'}$. This, together with the Lyapunov exponent of the system under consideration, determines a maximal interval of predictability within which the observational data x can be compressed. To preserve predictability beyond this interval, one is forced to update[17] his knowledge of the state of the system with a measurement of resolution k'. The phenomenon is well-known within chaos theory and shows up as a fundamental limitation to the predictability of such systems, a common exemple of which is weather.

Dynamical systems can generally be organized by considering the asymptotic of the string of measurements x with $N \to \infty$. At one end of the spectrum lie integrable systems, where k bits of knowledge of X_0 can be used all the way through to compute k bits of f_N. Those are systems where integration can be carried symbolically without an accumulation of errors. On the other side of this spectrum are chaotic systems, where $k + \lambda' N$ bits of X_0 are required to compute k bits of f_N. Similar classification schemes for dynamical systems that account for integrability and the appearance of chaos based on computational complexity have been proposed previously [22]. An algorithmic perspective on dynamical systems brings the possibility of considering other types of systems, where $k + g(N)$ bits of X_0 can be used to compute k bits of f_N, with $g(N)$ some *a priori* generic function.

3.4.2.2. *Thermodynamics and Statistical Mechanics*

Statistical mechanics posits the ergodicity of a complex dynamical system in order to obtain a partial, yet useful, description of its behaviour. This partial description is mostly understood to refer to the macroscopic description of a system displaying intractable microscopic descriptions. The generic approach is as follows. Starting again with an Hamiltonian and the associated phase space M, one first investigates the quantities conserved be the time evolution. By fixing those conserved quantities, one establishes constraints on phase space that restrict the accessible phase space to a bounded region. Properly defined, those constraints effectively decompose[18] the phase space into a family of submanifolds $F \subseteq M$ that are each preserved by time evolution. The ergodic hypothesis now posits that the curves X_t produced by an initial point $X_0 \in F$ under the time evolution are dense in each submanifold F such that the time average value of an observable $\mathcal{O} : M \to \mathbb{R}$, over such a curve is equal to the average of the same quantity over a uniform measure on

[17] It is here assumed that the Lyapunov exponent is constant and unique, which is not always the case.

[18] More precisely, these constraints generate a foliation of phase space that is invariant under the Hamiltonian flow.

each submanifold F.

$$\lim_{T\to\infty}\frac{1}{T}\int\limits_{t_0}^{t_0+T}\mathcal{O}(X_t)dt=\int\limits_{F_{X_0}\subseteq M}\mathcal{O}(X)d\mu(X),$$

for $d\mu$ the uniform measure over F. This uniform measure over the submanifolds F is often specified indirectly in terms of the Boltzmann weights of a state $X \in M$ over the submanifolds. With the above in mind, thermodynamics can be seen as the study of the interrelation of a relevant collection of macroscopic observables $\{\mathcal{O}_i\}$ expressing the change in the value of some observables in terms of the change in the value of the others. Such a thermodynamic description of a complex system is partial yet useful and relevant to the scale at which one would like to investigate the system.

Let us now concentrate on how this very generic picture of statistical mechanics and its relation to thermodynamics fits under our proposed definition of emergence. We first define the truncation $[F]_k$ to a finite resolution of a bounded[19] submanifold F as the set of strings which correspond to an encoding of a point in F to the prescribed k bits resolution. Then, positing the ergodicity of the system under study enables a direct reframing of statistical mechanics in terms of the ideas of this paper. Indeed, the postulated uniform measure on submanifolds of the phase space M amounts to postulating the corresponding microscopic states[20] in a submanifold to be equally likely under time evolution. In other words, for some large enough finite time interval τ, the sequence

$$x_N \equiv \langle\{[X^{(i)}_{n\tau}]_k\}_{1\leq i\leq \dim F,\, 0\leq n\leq N}\rangle,$$

is a typical sample of the truncated submanifold $[F]_k$. The lower bound on the time interval τ that needs to be satisfied for the above to hold is related to the Lyapunov exponent of the system. Indeed, such a bound corresponds to time intervals satisfying the converse of (3.10). In such a case, x_N is essentially an algorithmically random string. From this observation, it follows that for a sufficiently large time interval, one has that

$$K(x_N) = K([F]_k) + N\log(|[F]_k|), \tag{3.11}$$

[19] Here understood as meaning that the coordinate functions are all bounded. This guarantees a finite resolution in the truncation.

[20] Possibly with the exception of a measure zero set of states that are not relevant to the averaging of observables.

which indicates that the model $[F]_k$ for the string x_N is an algorithmic sufficient statistics.

The above discussion emphasized how, under the ergodic hypothesis, sufficient statistics are obtained by the specification of the associated decomposition into invariant submanifolds. A thermodynamical description of the system at equilibrium is in correspondence with such a decomposition of the phase space, provide the conserved quantities that define the submanifolds are taken to be the thermodynamical variables. Broadening the scope of the above, one can consider the possibility of an external interaction driving the system through different submanifold such that a complete description of the state X of the system can be made by the conjunction of a description of the entire decomposition of M into a collection of submanifolds $\{F\}$, an identification of the specific submanifold $F_X \ni X$ containing the state and an identification of the state within the submanifold F_X. Such a situation arrises when one considers a driven adiabatic thermodynamical process analyzed from the point of view of statistical mechanics. By first specifying the various decompositions of the phase space M associated to the thermodynamical variables, a shorter description of observational data x_N on this adiabatic process can be obtained as the state X_t at each measurement is confined to the intersection of the submanifolds associated to those thermodynamical variables, yielding a set of cardinality smaller than any of the individual $[F]_k$. The above illustrates how under the proposed definition, thermodynamics can be seen to be emerging from the exact description of complex ergodic systems. Of course, the discussion above is only exact inasmuch as the system size and measurement time are infinite so that the thermodynamic description become exact without the associated fluctuation theorems.

3.5. Conclusion

We proposed a definition of emergence casted in algorithmic information theory. This field has many times shown its usefulness to mathematically address mathematics, endowing it with a “meta” character. The absolute generality of AIT is a strength, but also a limitation and a vertigo of our proposal.

Intuitively, emergence is the appearance of unforeseen dynamics or properties exhibited by a complex system. In most discussions about emergence, the criteria of novelty highly depends upon the field: The aerodynamicist may be stunned by new patterns in fluid dynamics; The microbiologist, by new ways in which proteins could fold. In our proposed definition, emergence occurs in *"theory space"*: the thresholds of emergence are marked by the complexity of new models, which promote an over all shorter expression of the observed data. This is the essence of *understanding new structures*. Despite apparence, these models (sets of finite bit strings) are as general as they can be, since they are rooted in universal computation: Any "new pattern in fluid dynamics" or "new protein folding" named by a scientist is a computational process that can lead to an algorithmic model.

The development of our proposal was done through the "locally best models" of Kolmogorov's structure function. We called them the minimal partial models. In §3.3, we proved that:

(1) The data specifies almost everything about the minimal partial models;

(2) The magnitude of the drop measures the amount of "new understanding";

(3) Deeper minimal partial models almost specify the shallower ones.

We also extended the notions of coarse-grainings and boundary conditions, freeing them from any specific theory. In §3.4 we considered some applications to a toy model of a gas, dynamical systems and thermodynamics.

The absolute generality of algorithmic information theoretic methods come at the price of uncomputability. For instance, the shapes of Figure 3.10, in §3.4.1, are only conjectured, and will always be; No program can return the structure function of a piece of data x. A relaxation of the notion may be of interest in order to find concrete utility and applications in real life computation.

We shall stress that the concepts involved in the proof of Theorem 10 challenge the reconciliation between our mathematical proposal and the youth of our Universe. The deep models, namely those that occur at late drops of the structure function, are likely the result of programs that terminate after an unthinkably long computation... They have the largest finite running times among all programs no larger in size, so they solve the halting problem for shorter programs. This is the busy beaver regime. With a mere 14 billion years old, our Universe seems too young to accommodate such computations and this even holds if we take into account the parallel computation that occurs in the observable Universe. Indeed, the cube of Hubble's length over Plank's length is not even breakfast for the busy beaver!

Facing the realization that models witnessing drops of the structures are made of halting information, Vereshchagin and Shen [76] wrote "This looks like a failure. [...] [I]f we start with two old recordings, we may get the same information [about their minimal sufficient statistic], which is not what we expect from a restoration procedure. Of course, there is still a chance that some Ω-number [halting information] was recorded and therefore the restoration process indeed should provide the information about it, but this looks like a very special case that hardly should happen for any practical situation." Facing this, they suggest to consider models of more restricted classes or add some additional conditions and look for "strong models".

On the contrary, we think that that the minimal sufficient statistics of two recordings *should* share information, as they inevitably share a very common origin, which the model aims to capture. That this shared information is about the halting problem simply reflects the fact that their plausible common origin is the fruit of a very long computation, and not that the recording has anything to do with an Ω-number, or any representation of the halting problem. There is something very profound and unifying in the idea that emergent systems around us share some pieces of irreducible halting information, but this apparently requires an older Universe.

Let us open with philosophy of science. Let x be a string that encodes all scientific data ever recorded, together with meta data that may enhance its comprehension. Scientific theories aim at explaining x by grasping patterns in the data in order to reduce its redundancy. They are expressed in terms of models that distillate the structures from the apparently noisy boundary conditions. In an effort to tell apart theories or by pure experimental curiosity, the data x always increases, with the most recent bits reporting the latest experiments. Eventually, *better* theories arise, which may find important structures in apparent noise, or unify different models under the same umbrella. *This process of scientific investigation may be identified to the upper semi-computation of the structure function*. Indeed, through conjectures and guesses, theoreticians suggest models that can better explain the observations. These new best models can either be proven false by some eventually contradicting piece of data, or discarded on the basis of a simpler and fitter model, *i.e.*, closer to the graph of the actual structure function. But the new best models can never be proven right and this is the core idea of *fallibilism* in philosophy of science.

Acknowledgements

The authors are grateful to Philippe Allard Guérin, Charles Bennett, Gilles Brassard, Xavier Coiteux Roy and Pierre McKenzie for fruitful discussions. They also wish to thank the Institute for Quantum Optics and Quantum Information of Vienna, in particular, Marcus Huber's group, for support and discussions. The work of the authors is supported by Canada's Natural Sciences and Engineering Research Council (NSERC).

3.6. Appendix: Proof of Theorem 10

We suggest the reader to read the following article, Chapter 4, before looking at this proof, since the concepts involved in the proof are much more detailed in it.

Proof Referring to Chapter 4

From [75], one knows that $h_x(i) + i$ is logarithmically close to $\lambda_x(i)$, whose profile Λ_x satisfies $\Lambda_x \sim \mathcal{L}_x$. Hence, ε is chosen such that the drop in $h_x(i)+i$ forces the time profile $\mathcal{L}_x$ to drop. This ensures that the corresponding model, S, has a very slow S^*, otherwise the time profile would be contradicted by the two part description $D(S, i^x_S)$ (Recall that the second part of a two part description is fast). Hence, all the witnesses of the drops larger than ε of the structure function are indicative of algorithmic models almost full of halting information.

Proof without Reference to Chapter 4

Definitions Required for the Proof.

Definition 12. [3, Def 3] *The* Busy Beaver logical depth *of x with significance c is :*

$$Depth^B_c(x) \overset{df}{=} \min_{p,q}\{|q| \colon |p| \leq C(x) + c,\ \mathcal{U}(p) = x \text{ and } RT(p) \leq RT(q)\}.$$

Definition 13. [3, Def 7] *The* Set Sophistication *of x with significance c is :*

$$Soph^{Set}_c(x) \overset{df}{=} \min\{C(S) \colon x \in S \text{ and } C(S) + \log|S| \leq C(x) + c\}.$$

For reasonably small values of c, $\mathrm{Soph}^{\mathrm{Set}}_c(x)$ is the complexity α_M of the minimal sufficient statistics. As the values of c increase, $\mathrm{Soph}^{\mathrm{Set}}_c(x)$ will take the different values α_i of the complexities of the minimal partial models S_i.

Let the *upper graph* of a function f be $\{(n, m) \colon m \geq f(n)\}$. Let the distance between two points (n, m) and (n', m') be $\max(|n - n'|, |m - m'|)$.

Definition 14. [3, Def 4] *Two functions f and g are* ε–close *if the upper graphs of these functions are in a ε–neighbourhood of each other.*

Proof. We first want to show that $\mathrm{RT}(S^*_j) \geq B(\alpha_j - O(\log n))$. Let us first invoke the following theorem.

Theorem 15. [3, 77, Theorem 1 and Lemma 2] *For a fixed x of length n, the functions (of c) $\mathit{Depth}_c^{\mathsf{B}}(x)$ and $\mathit{Soph}_c^{\mathit{Set}}(x)$ are $O(\log n)$–close.*

Let i_j^x be the index of x in S_j. By giving S_j^* and i_j^x one obtains a two-part code Q for x of length $\alpha_j + \log|S_j| + O(1)$. Define c_j such that

$$\alpha_j + \log|S_j| = C(x) + c_j\,,$$

so $\mathrm{Soph}_{c_j}^{\mathrm{Set}}(x) = \alpha_j$ and $|Q| = C(x) + c_j + |a|$. Note that by definition of depth,

$$\mathrm{RT}(Q) \geq \mathsf{B}\Big(\mathrm{Depth}_{c_j+|a|}^{\mathsf{B}}(x)\Big)$$

and by the theorem,

$$\mathrm{Depth}_{c_j+|a|}^{\mathsf{B}}(x) \geq \mathrm{Soph}_{c_j+O(\log n)}^{\mathrm{Set}}(x) \pm O(\log n)\,.$$

If the drop at α_j is significative enough, namely, bigger then $O(\log n)$, than $\mathrm{Soph}_{c_j+O(\log n)}^{\mathrm{Set}}(x) = \alpha_j \pm O(\log n)$ (see figure?)

Hence,

$$\mathsf{B}^{-1}\left(\mathrm{RT}(Q)\right) \geq \alpha_j - O(\log n)\,.$$

Observe now that the running time of Q must be overwhelmed by the running time of S_j^*, since the second part of the program runs in time $O(|S_j|)$ which is negligible compared to $\mathsf{B}(\alpha_j - O(\log n))$. Therefore, $\mathrm{RT}(S_j^*) \geq \mathsf{B}(\alpha_j - O(\log n))$.

This means that S_j^* is such a slow program that from its running time, one can decide the halting problem for all programs of length $\leq \eta \overset{\mathrm{df}}{=} \alpha_j - O(\log n)$. The same holds for S_i^*, which can decide the halting problem for all programs of length $\leq \sigma \overset{\mathrm{df}}{=} \alpha_i - O(\log n)$.

Define ω_η as the number of programs of length $\leq \eta$ which halts. It is a condensed version of the halting oracle. This means that $K(\omega_\eta\,|\,S_j^*) = O(1)$ (or $K(\omega_\eta\,|\,S_j) = O(\log n)$) and since $K(\omega_\sigma\,|\,\omega_\eta) = O(1)$ and $K(\omega_\sigma) = \sigma$, then

$$\begin{aligned}
K(S_i\,|\,S_j) &\leq K(\omega_\sigma\,|\,S_j) + K(S_i\,|\,\omega_\sigma) + O(1)\\
&\leq K(S_i) - K(\omega_\sigma) + K(\omega_\sigma\,|\,S_i) + O(\log n)\\
&= \alpha_i - \sigma + O(\log n)\\
&= O(\log n)\,.
\end{aligned}$$

□

Chapitre 4

Relativity of Depth and Sophistication

Abstract. Logical depth and sophistication are two quantitative measures of the non-trivial organization of an object. Although apparently different, these measures have been proven equivalent, when the logical depth is renormalized by the busy beaver function. In this article, the measures are relativized to auxiliary information and re-compared to one another. The ability of auxiliary information to solve the halting problem introduces a distortion between the measures. Finally, similar to algorithmic complexity, sophistication and logical depth (renormalized) each offer a relation between their expression of (x, y), (x) and $(y|x)$.

4.1. Introduction

Around us are many objects that are neither completely trivial nor completely random. They conceal patterns and structures, buried under incidental disorganization. As Bennett [16] coins it, they "contain internal evidence of a nontrivial causal history". Such objects are difficult to model and to explain, yet, *interesting*. And interesting itself is the task of formalizing mathematically this very notion. Computability theory has led to the development of algorithmic information theory (AIT) and computational complexity theory, two domains in which formal notions for this "interestingness" have been casted.

Embedded in AIT is the approach of nonprobabilistic statistics, proposed by Kolmogorov [52] in the mid 70's, which attempts to distil the "concealed patterns and structures" from the apparent "incidental disorganization". As in probabilistic statistics, the mission of this approach is to find the most plausible model that supports the object. Such a model is identified to the simplest one that entails a nearly shortest description of the object in two parts. The first part describes the model (structures and patterns) and the second part is a canonical specification of the precise object among all of which are consistent with the model (incidental randomness). Kolmogorov pointed out that the description length of such a model is a value of particular interest. Koppel [53] (indirectly) referred to this quantity as the *sophistication* of the object, a first notion of interestingness.

Unlike probabilistic statistics, however, an individual object is considered, dismissing anything else "it could have been". It is not hypothesized to be drawn from some *unexplained probabilistic* process; instead, it is supposed to have originated from an *unknown computable* process[1]. This assumption goes hand in hand with the physical Church-Turing's thesis, namely, the belief that physical processes can be simulated with arbitrary accuracy by a universal computer. Indeed if the object comes from "around us" it has originated from an unknown physical process, whence the aforementioned assumption.

The other approach to quantify interestingness is from a radically different angle, incorporating ideas from computational complexity theory to AIT. In the seminal paper [51] in which he defines algorithmic complexity, Kolmogorov concludes by mentioning the "existence of cases in which an object permitting a very simple program, *i.e.*, with very small complexity $K(x)$, can be restored by short programs only as the result of thoroughly unreal duration". He then writes of his intention of further studying the topic, but he published nothing later on the subject. More than twenty years later, in the late 80's, Bennett carried the torch. The most plausible causal histories of an object lie in the shortest computable descriptions. If all those descriptions entail a lengthy computation, this signifies a difficult deductive path and hence non-triviality of the object. Its *Logical depth* is then the running time of its most plausible computable description.

Although many people [53, 4, 6, 10] had observed connections between (variants of) sophistication and logical depth, it is only recently that they have been identified [3] as the same quantity, when logical depth is renormalized to map the "thoroughly unreal duration" back into a number comparable to a program length (*e.g.*, the length of a model description). In this paper, I analyse further those two apparently different — but in fact equivalent — approaches to measure the buried structures of an interesting object.

Algorithmic complexity satisfies the chain rule, Eq. (4.1), which *connects* the complexity of a pair (x,y) the complexity of x and the complexity of y relative to x. *The goal of*

[1] This justifies the name "algorithmic statistics" also used as a synonym of nonprobabilistic statistics.

this paper is to investigate whether sophistication and depth also exhibit such a connection between (x,y), (x) *and* $(y|x)$. The main exploration then regards the *relativity* of depth and sophistication, namely, how the concepts change when the universal computer is supplemented with auxiliary information. I show that when both are relativized, sophistication no longer amounts to the renormalized logical depth (§4.3). Their difference is shown to be a function of the difficulty to materialize the halting information of the auxiliary string (§4.4). I then reach the goal: I demonstrate that the depth (again, the renormalized version) of a pair of objects (x,y) can be expressed as the maximum between the depth of x and the depth of y relative to x; sophistication of a pair admits a similar, yet distorted relation (§4.5). Finally, I revisit the so-called antistochastic strings from running time considerations (§4.7).

4.2. Preliminaries

Established notions of AIT and nonprobabilistic statistics, as well as elementary reformulations and generalizations are presented in this section. For attributions and more details, see Refs. [59, 72].

Notation

I denote $\mathbb{N} = \{0,1,2,\ldots\}$ and $\{0,1\}^* = \{\epsilon,0,1,00,\ldots\}$. I refer to finite bit strings simply as "strings". The first i bits of a (finite or infinite) string x is denoted $x_{[i]}$. The length of a string x and the cardinality of a set S are denoted $|x|$ and $|S|$; the context will distinguish the meaning. A quantity Q may depend on some parameter n. The quantity $O(g(n))$ $[\Omega(g(n))]$ denotes a positive function eventually upper bounded [lower bounded] by $cg(n)$, where c is a constant. I write $Q \asymp f(n)$, $Q \preccurlyeq f(n)$ and $Q \succcurlyeq f(n)$ if, respectively, $Q(n) - f(n) = \pm O(1)$, $Q(n) \leq f(n) + O(1)$ and $Q(n) \geq f(n) - O(1)$. I write $Q \sim f(n)$, $Q \lesssim f(n)$ and $Q \gtrsim f(n)$ if, respectively, $Q(n) - f(n) = \pm O(\log n)$, $Q(n) \leq f(n) + O(\log n)$ and $Q(n) \geq f(n) - O(\log n)$.

4.2.1. Algorithmic Complexity

The question of whether — and if so how — one can robustly represent objects "around us" digitally (*i.e.*, using a finite alphabet) is not simple. It falls in the realm of philosophy of science, not that of coding theory. For a discussion on the topic, see

Ref. [11]. Nonetheless, digital objects can easily be encoded in strings, thereby restricting the theory to the latter. The algorithmic complexity $K(x)$ of a string x is the length of the shortest program to compute x on a universal computer. For a meaningful definition, a model of computation and a universal computer within the model need to be fixed. However, from the ability of universal computers to simulate one another and, by the Church-Turing thesis, to simulate any computable process, the algorithmic complexity of a string is independent of the fixed universal computer, up to an additive constant. In this sense, the algorithmic complexity can then be viewed as a universal and absolute quantity of information — or randomness — in a string.

Chaitin [24] defines a similar model in which, the universal computer $\mathcal{U}$ is fixed to be a *self-delimiting* Turing machine, *i.e.*, it has a read-only one-way input tape and some work tapes. When the computation begins, a program p occupies the input tape and an auxiliary string z occupies a designated work tape. The computation succeeds only if the machine reaches a halting state while its read head is scanning the rightmost bit of p, but no further. This forces the program to contain within itself the information about its own length. A successful computation is denoted by $\searrow$ and $\mathcal{U}(p,z)$ is then defined to be the string displayed on the work tape at halting. Self-delimitation ensures that for any z the set $\{q\colon \mathcal{U}(q,z)\searrow\}$ is a *prefix-free* set of strings, namely, no member of which is a prefix of another. When no auxiliary information is provided, z is simply set to ϵ, and $\mathcal{U}(p,\epsilon)$ is abbreviated to $\mathcal{U}(p)$.

The (prefix) *algorithmic complexity* is defined with respect to the above universal computer $\mathcal{U}$ as

$$K(x) \overset{\mathrm{df}}{=} \min_p \{|p|\colon \mathcal{U}(p) = x\},$$

and its *conditional* counterpart as

$$K(x|z) \overset{\mathrm{df}}{=} \min_p \{|p|\colon \mathcal{U}(p,z) = x\}.$$

Multiple strings can be encoded into a single one via a computable bijection $(x_1, x_2, \dots, x_n) \mapsto \langle x_1, x_2, \dots x_n \rangle$ uniformly defined for any n. The complexity of multiple strings is thus naturally defined as $K(\langle x_1, x_2, \dots, x_n \rangle)$.

Let x^* and $(x|z)^*$ be the[2] shortest programs that computes x with ϵ and with z as auxiliary information, respectively.

Remark 1. *Observe that*

$$\begin{tikzcd} & & x \\ x^* \arrow[r] & \boxed{O(1)} \arrow[ur] \arrow[dr] & \\ & & K(x) \end{tikzcd} \quad \textit{and} \quad \begin{tikzcd} x \arrow[dr] & & \\ & \boxed{O(1)} \arrow[r] & x^* \\ K(x) \arrow[ur] & & \end{tikzcd},$$

where the diagrams represent that the output(s) can be computed from the input(s) and a $O(1)$ advice. Indeed, $K(x)$ and x can be computed from x^* by measuring its length before executing it. And x^* can be determined by a parallel execution of programs of length $K(x)$, until x is produced.

A very important relation is the *chain rule*,

$$K(x,y) \asymp K(x) + K(y|x^*), \tag{4.1}$$

as it entails a symmetric notion of *mutual information*, so defined as

$$I(x:y) \overset{\text{df}}{=} K(y) - K(y|x^*).$$

The "$\preccurlyeq$" side of Equation (4.1) is easily observed, as one way to compute $\langle x,y\rangle$ is to copy and then execute x^*, which can then serve as an auxiliary string to $(y|x^*)^*$. At this stage, $\langle x,y\rangle$ can be computed. The "$\succcurlyeq$" side, harder to prove, states that the previous procedure to compute $\langle x,y\rangle$ is nearly optimal in terms of program length.

Observe that by the information equivalence of x^* and $\langle x, K(x)\rangle$, Remark 1, $K(y|x^*) \asymp K(y|x,K(x))$. This is convenient to write the relativized chain rule as

$$K(x,y|z) \asymp K(x|z) + K(y|x,K(x|z)).$$

Halting Information

To determine whether, for a given p, $\mathcal{U}(p)$ is a halting computation or not is an undecidable task. The *halting problem* is perhaps the most famous of computability theory. It can perfectly be framed in AIT, and even, better quantified.

As suggested by Turing [74], the halting problem can be encoded into bits. The most straightforward way of doing so is to define the infinite string $\mathcal{H}$ whose i-th bit is 1 if and only if the i-th program, in lexicographic order, halts. I denote $\mathcal{H}^{\leq j}$ the first $2^{j+1}-1$

[2] In the case of multiple programs of minimal length, the fastest trumps.

bits of $\mathcal{H}$, which encode the solution to the halting problem for all programs of length $\leq j$. Such a representation of the halting problem is highly redundant, since the same information can be given in much fewer bits. In fact, together with j, the number ω_j of programs of length $\leq j$ that halt suffices, because one can recover $\mathcal{H}^{\leq j}$ by running all programs no longer than j in parallel until ω_j of them have halted.

A more elaborate way of encoding the halting problem is through Chaitin's *halting probability* [24] defined as

$$\Omega = \sum_{p:\mathcal{U}(p)\searrow} 2^{-|p|}.$$

Since the set of halting program is prefix-free, Kraft inequality implies that the sum converges to a number smaller than 1. If a program is given to the reference machine $\mathcal{U}$ with bits picked at random, then the probability that the computation ever halts is Ω. The first j bits of Ω, denoted $\Omega_{[j]}$, can be used to compute $\mathcal{H}^{\leq j}$,

$$\Omega_{[j]} \longrightarrow \boxed{O(1)} \longrightarrow \mathcal{H}^{\leq j}.$$

This is done by running all programs in a dovetailed fashion, and adding $2^{-|p|}$ to a sum M (initially set to 0) whenever a program p halts. When the first j bits of the sum stabilize to the first j bits of Ω, *i.e.*, $M_{[j]} = \Omega_{[j]}$, then no program of length $\leq j$ will ever halt, since such an additional contribution to the sum would contradict the value of Ω. This process is said to *lower semi-compute* Ω, since it always returns smaller numbers than Ω and they converge to it in the limit of infinite time.

Ω is an example of an *incomressible* string, namely that $K(\Omega_{[j]}) \succcurlyeq j$. This can be proved from a *Berry paradox* argument: the ability of $\Omega_{[j]}$ to compute $\mathcal{H}^{\leq j}$ also endows it with the ability to produce ζ, the first string in lexicographic order with complexity $> j$. However, such a computation of ζ from $\Omega_{[j]}$ is only consistent if $K(\Omega_{[j]}) \succcurlyeq j$. Moreover, as any string of length j, $\Omega_{[j]}$ has (prefix) complexity $\preccurlyeq j + K(j)$. Hence,

$$j \preccurlyeq K(\Omega_{[j]}) \preccurlyeq j + K(j).$$

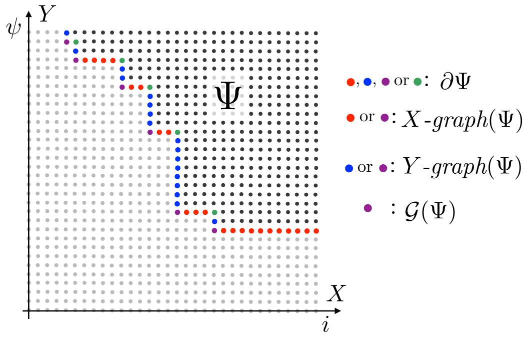

ψ
Y
Ψ
●, ●, ● or ●: $\partial\Psi$
● or ●: X-*graph*(Ψ)
● or ●: Y-*graph*(Ψ)
● : $\mathcal{G}(\Psi)$
X
i

Fig. 4.1. The anatomy of a profile $\Psi \subseteq \mathbb{N}^2$.

4.2.2. Nonprobabilistic Statistics

Before overviewing the algorithmic treatment of statistics, I introduce elementary concepts and notations about subsets of $\mathbb{N}^2$. They will be useful for illustrational purposes, conciseness of notation and most importantly to unify different definitions under the same umbrella.

The Help of $\mathbb{N}^2$

A set $\Psi \subseteq \mathbb{N}^2$ is *upwards closed* [resp. *rightwards closed*] if

$$(i,\psi) \in \Psi \implies \forall k,\ (i,\psi+k) \in \Psi \text{ [resp. } (i+k,\psi) \in \Psi].$$

A *profile* is an upwards and rightwards closed subset of $\mathbb{N}^2$. The L^∞-metric endows $\mathbb{N}^2$ with a distance. The *distance* between (a_1,a_2) and (b_1,b_2) is given by $\max(|a_1-b_1|,|a_2-b_2|)$.

Let Ψ be a profile. Its *boundary* $\partial\Psi$ is the subset at distance unity of some point outside of Ψ, *i.e*, each point in $\partial\Psi$ has at least one of its 8 neighbours outside of Ψ. The X-*graph* of Ψ is

$$X\text{-}graph(\Psi) \overset{\text{df}}{=} \{(i,\psi) \in \partial\Psi : (i,\psi-1) \notin \Psi\}.$$

It is the graph of some function $\psi(i)$ represented as usual by the Y versus X axes. The Y-*graph* of Ψ is analogously defined as

$$Y\text{-}graph(\Psi) \overset{\text{df}}{=} \{(i,\psi) \in \partial\Psi : (i-1,\psi) \notin \Psi\},$$

and is the graph of some function $i(\psi)$ unusually represented by the X versus Y axes. See Figure 4.1.

Remark 2. *(Let Ψ be upwards and rightwards closed.) Both functions $\psi(i)$ and $i(\psi)$, represented respectively by the X-*graph *and the Y-*graph, *are non-increasing. These functions are in general noninvertible, but they are as close as they can get from being each other's inverse, specifically,*

$$\psi(i') = \psi' \implies i(\psi') \leq i' \qquad \textit{and} \qquad i(\psi') = i' \implies \psi(i') \leq \psi'.$$

A set $G \subseteq \mathbb{N}^2$ is said to *generate* Ψ if the upwards and rightwards closure of G gives Ψ. Such a closure is understood to be $\{(i', \psi') \in \mathbb{N}^2 : \exists (i, \psi) \in G \ i \leq i' \text{ and } \psi \leq \psi'\}$. Of a particular interest is the minimal such set. The *Generator set* of Ψ is defined as

$$\mathcal{G}(\Psi) \overset{\text{df}}{=} X\textit{-graph}(\Psi) \ \cap \ Y\textit{-graph}(\Psi).$$

It corresponds to the convex corners of Ψ, namely, the corners that have more neighbours outside than inside Ψ.

The *sum* of two profiles Ψ and Φ is defined as

$$\Psi + \Phi \overset{\text{df}}{=} \{(i, \psi + \phi) : (i, \psi) \in \Psi \text{ and } (i, \phi) \in \Phi\}.$$

The *ε-neighbourhood* of Φ includes all points at a distance $\leq \varepsilon$ of each of its points, hence enlarging the boundary. Ψ is *ε-close* to Φ if it is contained in an *ε-neighbourhood* of Φ.

Remark 3. *(Let Ψ and Φ be upwards and rightwards closed.) Ψ is ε-close to Φ*

*(i) if and only if $\mathcal{G}(\Psi)$ it is contained in an ε-*neighbourhood *of Φ*

(ii) if and only if $\psi(i) + \varepsilon \geq \phi(i + \varepsilon)$,

where ψ and ϕ are the functions represented by the respective X-graphs.

I denote $\Psi \asymp \Phi$ or $\Psi \sim \Phi$ if Ψ and Φ are both $O(1)$-close or $O(\log n)$-close to one another, respectively. Those relations find their usefulness in the *two-dimensionality of the approximation*, which cannot be expressed so concisely, for example, by the X-graphs.

Quantifying "Good" Models

For a review of the field of nonprobabilistic statistics, see Ref. [76].

A finite set S that contains a string x is an *algorithmic statistic* of x. It is also called a *model* of x, since it puts together strings that share common properties with x, precisely those that define S. Opposing qualities are expected of a good model. On the one hand, the model should be simple, tending to minimize $K(S)$. The latter is the length of the shortest program that computes an encoding of the lexicographical ordering of the elements of S and halts. On the other hand, the canonical description of x *via the model* should also be minimized. In the case of finite sets as models, such a description amounts to describing first S and then specifying $x \in S$ by some canonical encoding, for instance, by giving its index i_S^x in a lexicographical ordering of the elements of S.

More precisely, each model $S \ni x$ entails a *two-part description* of x. The first part consists of describing the model by its shortest program S^* (of length $K(S)$) and the second part singles out x in S, thanks to its index i_S^x (of length $\log|S|$). This second part is known as the *data-to-model code*, but really, it should be called the *model-to-data code*. This means that

$$D(S^*, i_S^x) \overset{\text{df}}{=} \alpha S^* i_S^x$$

is a self-delimiting program that computes x, where the prefix α is a fixed program (of length $O(1)$) which ensures the correct execution of the two-part description. Note that the second part of the code does not need any additional prefix for self-delimitation, since its length $|i_S^x| = \lceil \log|S| \rceil$ can be computed (by α) from S^*. The length of the two-part description is therefore given by

$$|D(S^*, i_S^x)| = K(S) + \log|S| + |\alpha|.$$

The tradeoff between the simplicity of the model and the length of its corresponding two-part description can be expressed by a profile on $\mathbb{N}^2$: for each $S \ni x$, a dot can be marked at the coordinate $(K(S), K(S) + \log|S| + |\alpha|)$. The upwards and rightwards closure of those dots yields what I call the *description profile*,

$$\Lambda_x = \{(i, \lambda) \colon \exists S \ni x,\ i \leq K(S) \text{ and } K(S) + \log|S| + |\alpha| \leq \lambda\}.$$

The *X-graph* of Λ_x represents what is known [75] as the *constrained minimum description length function*

$$\lambda_x(i) = \min_{S \ni x}\{K(S) + \log|S| + |\alpha| \colon K(S) \leq i\}.$$

For i large enough, $\lambda_x(i)$ reaches values close to $K(x)$. In the worse case, this is achieved for $i \asymp K(x)$ as witnessed by the model $\{x\}$. A model S that entails a two-part description essentially as short as the shortest program is called *sufficient*. Kolmogorov pointed out that a sufficient model S_0 of *minimal* complexity describes all the structure of x, or in Vitányi's words [77], its "meaningful information", but not more. The remaining information $i^x_{S_0}$ is the incidental or random part of x. The complexity $K(S_0)$ of a minimal sufficient statistics is now known as the *sophistication* of x. For a precise definition, one needs to clarify what is meant by "reaches values close to $K(x)$", which introduces a resolution parameter c,

$$\mathrm{Soph}_c(x) \overset{\mathrm{df}}{=} \min_{S \ni x}\{K(S)\colon\ K(S) + \log|S| + |\alpha| \leq K(x) + c\}.$$

Although the sophistication of a string x is intuitively thought to be the value of $\mathrm{Soph}_c(x)$ for a resolution c as small as possible, it is meaningful to view $\mathrm{Soph}_c(x)$ as a function of c since it allows to connect sophistication with the description profile Λ_x. First, one translates Λ_x down on the Y axis by $K(x)$ to define

$$\begin{aligned}\Delta_x &\overset{\mathrm{df}}{=} \Lambda_x - (0, K(x)) \\ &= \{(i,c)\colon\ \exists S \ni x,\ K(S) \leq i \text{ and } K(S) + \log|S| + |\alpha| \leq K(x) + c\}.\end{aligned}$$

The Y-graph of Δ_x is obtained by minimizing the first coordinate, with the second coordinate fixed, yielding $(\mathrm{Soph}_c(x), c)$.

Robustness of the Method

The method used to arrive at a definition of sophistication may appear somewhat arbitrary. Among the different model-selection principles, why minimizing the two-part description? And why imposing finite sets as a model class? Each of these issues have been specifically addressed and the method shows robustness since different model-selection principles and different model classes yields essentially the same measure of sophistication.

In the method presented here, the quality opposed to the simplicity of the model was the minimality of the two-part description, known as the *minimum description length principle*. The trade-off between those qualities is expressed by the function $\lambda_x(i)$ from which sophistication was read out. Another quality of a model that opposes its simplicity is guided by the *maximum likelihood principle*, which favours the models with as few elements as possible. This trade-off is displayed by the *constrained maximum likelihood* *function*,

$$h_x(i) = \min_{S \ni x}\{\log|S| \colon K(S) \leq i\}.$$

This is Kolmogorov's original [52] *structure function*. Another principle is to minimize the *randomness deficiency*, valuing models S in which x is most typical. This defines the function

$$\beta_x(i) = \min_{S \ni x}\{\log|S| - K(x|S) \colon K(S) \leq i\},$$

since the lack of typicality is measured by how far from the data-to-model code is the shortest program for computing x given S.

Importantly, Vereshchagin and Vitányi [75] showed that the three functions λ_x, h_x and β_x encode the same information, since they are all connected to each other by affine transformations (within logarithmic precision). In particular, the minimal value at which $\lambda_x(i)$ reaches close to $K(x)$, that is, the sophistication, can be defined alternatively from the maximum likelihood or the randomness deficiency principles. In this paper, the attention is restricted to λ_x, or more specifically, to its corresponging description profile Λ_x.

The other critique that can be formulated about the path used to define sophistication is the lack of generality of finite sets as a model class. In fact, some people [44, 43] have generalized the model class to computable probability distributions, possibly infinite. The complexity of the model then becomes that of the distribution and the length of the data-to-model code is then given by the Shannon-Fano code. The constrained minimum description length function, analogous to λ_x, is then expressed in terms of these quantities, and again the value at which the function reaches close to $K(x)$ is identified. Gell-Mann and Lloyd called it *effective complexity* [44]. Yet one more model class possibly even more general is given by total functions[3], where again, two part-descriptions are analogously defined.

Vitányi [77] showed that whether the model class is fixed to finite sets, computable distributions or total functions, the respective description profiles would be close to one another, underlining again the robustness of sophistication, and the sufficiency of finite sets as model class.

[3] In fact, the term sophistication was coined by Koppel as he was grasping the idea through total functions as a model class.

Finally, a very important result of algorithmic statistics states that the description profile Λ_x can essentially take all possible shapes.

Theorem 4 (All shapes are possible [75])**.** *Let $k \leq n$. Let $\mathcal{G}$ be some set of points that generates a profile $\mathcal{T}$ by upwards and rightwards closure in such a way that $(0,n) \in T$ and $(k,k) \in \partial T$. Then there exists a string x of complexity $k + O(\log n) + K(\mathcal{G})$ and length $n + O(\log n) + K(\mathcal{G})$ whose description profile Λ_x is $O(\log n) + K(\mathcal{G})$-close to $\mathcal{T}$.*

4.2.3. Logical Depth and Time-Bounded Complexity

One of the most beautiful surprises of algorithmic statistics is that its core concepts are directly related to running-time considerations.

Hereinafter, $\mathrm{RT}(p)$ stands for the *running time* of p, which is the number of computation steps that $\mathcal{U}$ executes on input p before reaching a halting state. If the computation uses auxiliary information z, then I denote $\mathrm{RT}(p,z)$ the running time of the computation $\mathcal{U}(p,z)$.

An object x is *deep* if most of its algorithmic probability corresponds to slow computations. The gist of this idea is captured by Bennett's second tentative definition [16] of *logical depth*, with significance parameter c:

$$\mathrm{Depth}_c(x) \overset{\mathrm{df}}{=} \min_{p\colon \mathcal{U}(p)=x} \{\mathrm{RT}(p)\colon |p| \leq K(x) + c\}.$$

Running times can be very large, especially when interested by the deepest strings of a fixed length. The *inverse busy beaver function* renormalizes those astronomical running times back into numbers of size comparable to program length.

Definition 5. *The* busy beaver *is a function $B\colon \mathbb{N} \to \mathbb{N}$ defined by*

$$B(n) \overset{df}{=} \max\ \{RT(p)\colon \mathcal{U}(p) \searrow \ and\ |p| \leq n\}.$$

It is the maximal finite running time of a program of n bits or less. Its inverse $\mathsf{B}^{-1}(N)$ is then defined as the length of the shortest program that eventually halts after at least N steps:

$$\mathsf{B}^{-1}(N) = \min\{|p|\colon \mathrm{RT}(p) \geq N \text{ (but finite)}\}.$$

As a convenient shortcut, one can measure time right away in busy beaver units by defining the *busy running time* $\tau(p)$ of a program p as

$$\tau(p) \overset{\mathrm{df}}{=} \mathsf{B}^{-1}(\mathrm{RT}(p)).$$

Deploying this definition, if p has a busy running time $\tau(p) = d$, it means that there is a program of length d, but none of length less than d, that halts after p.

Definition 6. *The* busy beaver depth *of x, at significance level c is defined here like in Ref.* [3], *but with prefix instead of plain complexity:*

$$Depth^B_c(x) \overset{df}{=} \min_{p:\,\mathcal{U}(p)=x} \{\tau(p) \colon |p| \le K(x) + c\}.$$

It amounts to the inverse busy beaver of the logical depth[4].

Related to logical depth is the concept of *time-bounded complexity*,

$$K^t(x) \overset{\mathrm{df}}{=} \min_{p:\,\mathcal{U}(p)=x} \{|p| \colon \mathrm{RT}(p) \le t\}.$$

The notion was already mentioned in the conclusions of Kolmogorov's seminal paper [51], as a proposed tool to "study the relationship between the necessary complexity of a program and its permissible difficulty t". The quoted relationship can be explored through the *time profile* $\mathcal{L}_x$, generated by the coordinates $(\tau(p), |p|)$ for each program p that computes x. Written differently,

$$\mathcal{L}_x = \{(i, \ell) \colon \exists p\ \mathcal{U}(p) = x,\ |p| \le \ell \text{ and } \tau(p) \le i\}.$$

Observe that

$$\mathcal{L}_x = \{(i, \ell) \colon K^{B(i)}(x) \le \ell\}, \tag{4.2}$$

so $(i, K^{B(i)})$ is the *X-graph* of $\mathcal{L}_x$. By a process analogous to the reading out of sophistication from the description profile Λ_x, the busy beaver depth can be expressed from the time profile $\mathcal{L}_x$. To do so, define

$$\mathcal{D}_x \quad \overset{\mathrm{df}}{=} \quad \mathcal{L}_x - (0, K(x))$$

[4] The definition of logical depth on which Bennett settled in Ref.[16] imposes the condition $K(p) \ge |p| - c$ instead of $|p| \le K(x) + c$. It has been shown [3] that in the plain complexity setting, the inverse busy beaver renormalization of such a definition of logical depth is $O(1)$ close to the plain complexity counterpart of Def. 6, up to $O(1)$ precision also in the significance parameter.

$$= \{(i,c)\colon \exists p,\ \mathcal{U}(p)=x,\ \tau(p)\le i \text{ and } |p|\le K(x)+c\}.$$

The Y-graph is obtained by minimizing the first coordinate, with second coordinate fixed, yielding $(\mathrm{Depth}_c^{\mathsf{B}}(x), c)$.

The following remarkable result connects the description and time profiles, and so sophistication and depth.

Theorem 7 ([3, 10])**.** *For all x,*

$$\mathcal{L}_x \sim \Lambda_x \qquad \textit{and so} \qquad \mathcal{D}_x \sim \Delta_x.$$

4.2.4. Definitions Relativized

The previous definitions capture properties of a fixed bit string x. The same definitions also hold if one reads x as an encoding $\langle x', y'\rangle$ of a pair of strings. The main intention of this paper is to study the properties of description and time profiles when they are relativized[5] by some auxiliary information z. Here, I straightforwardly extend the definitions to a conditional counterpart.

The *conditional description and time profiles* are respectively

$$\begin{aligned}\Lambda_{y|z} &= \{(i,\lambda)\colon \exists S \ni y\ i \le K(S\,|\,z) \text{ and } K(S\,|\,z)+\log|S|{+}|\alpha| \le \lambda\}\\ \mathcal{L}_{y|z} &= \{(i,\ell)\colon \exists p\ \mathcal{U}(p,z)=x,\ |p|\le \ell \text{ and } \tau(p,z)\le i\}.\end{aligned}$$

Sophistication, busy beaver depth and time-bounded complexity also have a straightforward conditional analogues:

$$\begin{aligned}\mathrm{Soph}_c(y\,|\,z) &\overset{\text{df}}{=} \min_{S\ni y}\{K(S\,|\,z)\colon K(S\,|\,z)+\log|S|+|\alpha|\le K(y\,|\,z)+c\},\\ \mathrm{Depth}_c^{\mathsf{B}}(y\,|\,z) &\overset{\text{df}}{=} \min_{p\colon \mathcal{U}(p,z)=y}\{\tau(p,z)\colon |p|\le K(y\,|\,z)+c\},\\ K^{\mathsf{B}(i)}(y\,|\,z) &\overset{\text{df}}{=} \min_{p\colon \mathcal{U}(p,z)=y}\{|p|\colon \mathrm{RT}(p,z)\le \mathsf{B}(i)\}.\end{aligned}$$

One can again translate profiles

$$\Delta_{y|z} \overset{\text{df}}{=} \Lambda_{y|z} - (0, K(y\,|\,z)) \qquad \text{and} \qquad \mathcal{D}_{y|z} \overset{\text{df}}{=} \mathcal{L}_{y|z} - (0, K(y\,|\,z))$$

[5] In this paper, I use "relativized by" in the same sense as "conditional to".

and verify that the definitions are consistent with

$$\begin{aligned} Y\text{-}graph(\Delta_{y|z}) &= (\mathrm{Soph}_c(y\,|\,z), c) \\ Y\text{-}graph(\mathcal{D}_{y|z}) &= (\mathrm{Depth}^{\mathsf{B}}_c(y\,|\,z), c) \\ X\text{-}graph(\mathcal{L}_{y|z}) &= (i, K^{\mathsf{B}(i)}(y\,|\,z)) \\ X\text{-}graph(\Lambda_{y|z}) &= (i, \lambda_{y|z}(i)). \end{aligned}$$

4.3. Chain Rules for Profiles

One of the most important relations in AIT is the chain rule, eq. (4.1), for algorithmic complexity. Without it, the “IT” in “AIT” would be a misnomer, because like in Shannon’s theory of information, the chain rule is precisely what entails a symmetric measure of information. A chain rule for time and description profiles would make it possible to express depth and sophistication of a pairs in terms of their single string and conditional version.

4.3.1. A Chain Rule for Time Profiles

In this section, I show that the chain rule is carried over by the time profiles within logarithmic resolution, namely, that the following holds

$$\mathcal{L}_{x,y} \sim \mathcal{L}_x + \mathcal{L}_{y|x}. \tag{4.3}$$

The above relation accounts for two “profile inequalities”. Each of which is treated independently in Proposition 8 and Proposition 9, because they hold with different error bounds. In both propositions, the strategy is the same: I follow the lines of Longpré’s analysis [60] of the chain rule for time-bounded complexity, but where time is measured in the busy beaver scale and where programs are required to be self-delimited.

Proposition 8. *For all strings x and y of length $\leq n$ and for all $i \geq B^{-1}(n)$,*

$$K^{B(i')}(x,y) \preccurlyeq K^{B(i)}(x) + K^{B(i)}(y\,|\,x),$$

where $i' = i + O(1)$.

Proof. Let p and q be the respective witnesses of $K^{\mathsf{B}(i)}(x)$ and $K^{\mathsf{B}(i)}(y\,|\,x)$. Then rpq is a self-delimiting program for $\langle x,y\rangle$, where r is a constant-size routine that implements the following. First, p is executed, producing x, which is copied before being given as a ressource to q. Thereupon, q is executed, yielding y, and the pair $\langle x,y\rangle$ is computed. The running time of the whole computation is $2\mathsf{B}(i)+O(n) \preccurlyeq O(\mathsf{B}(i)) \preccurlyeq \mathsf{B}(i+O(1))$. □

Proposition 9. *For all strings x and y of length $\leq n$ and for all $i \geq B^{-1}(n)$,*

$$K^{B(i')}(x)+K^{B(i')}(y\,|\,x) \preccurlyeq K^{B(i)}(x,y)+2K^{B(i)}(m,l),$$

where $i' = i+O(1)$, $K^{B(i)}(x,y)=m$ and $l \leq m$ is to be determined.

Proof. Define

$$A^i = \{\langle x',y'\rangle\colon K^{\mathsf{B}(i)}(x',y') \leq m\} \text{ and } \qquad A^i_x = \{y'\colon K^{\mathsf{B}(i)}(x,y') \leq m\},$$

which contain $\langle x,y\rangle$ and y, respectively. A program for y given x is to enumerate A^i_x and to give its enumeration number i^y. A^i_x can be enumerated if x and m are known. Note that $\mathsf{B}(i)$ is not required, since the enumeration can be done in a parallel fashion until the i^y-th element has been enumerated. This takes a maximum of

$$O\Big(2^{m+1}(1+2+\cdots+\mathsf{B}(i))\Big) = O\Big(2^m\mathsf{B}(i)^2\Big)$$

steps of computation, namely, enough for each program of length $\leq m$ to be executed (parallel fashion) for $\mathsf{B}(i)$ steps. The exponential factor in $O(2^m\mathsf{B}(i)^2)$, which can be furthermore bounded by $\leq O(2^{O(\mathsf{B}(i))}\mathsf{B}(i)^2)$, seems like bad news (it would be for concrete computations). But when compared to $\mathsf{B}(i+O(1))$, it is safely ignored. In fact, for any computable function f,

$$\mathsf{B}(i+O(1)) \geq f(\mathsf{B}(i)),$$

because the program of length i that runs for $\mathsf{B}(i)$ and the program of length $O(1)$ that computes the function $f(\cdot)$, can be merged into a program of length $i+O(1)$ that runs for $f(\mathsf{B}(i))$.

Let $l \equiv \lceil \log|A^i_x|\rceil$ be the number of bits of i^y. Self-delimitation of the program for y given x imposes that l must be known in advance. Hence, the program considered requires $K^{\mathsf{B}(i)}(m,l)$ bits to compute m and l (in time $\leq \mathsf{B}(i)$), and l bits to give the enumeration number of y. Any $O(n)$ execution times are absorbed in $\mathsf{B}(i+O(1))$, so

$$K^{\mathsf{B}(i+O(1))}(y\,|\,x) \preccurlyeq l + K^{\mathsf{B}(i)}(m,l). \tag{4.4}$$

Define now

$$B^i = \{x' \colon \log|A^i_{x'}| > l - 1\},$$

which contains x. If m and l are given, B^i can be enumerated by enumerating A^i (thanks to m), and when for a given x' the subset $A^i_{x'}$ contains more than 2^{l-1} elements, x' is added to B^i. A possible program for x is thus given by the enumeration number of x in B^i. The enumeration of A^i will be completed in time $\leq \mathsf{B}(i + O(1))$, after which x is guaranteed to have appeared in the B^i list. Note that

$$|A^i| = \sum_{x'} |A^i_{x'}| \geq \sum_{x' \in B^i} |A^i_{x'}| \geq \sum_{x' \in B^i} 2^{l-1} = |B^i| 2^{l-1}.$$

Since $|A^i| < 2^{m+1}$,

$$\log|B^i| \preccurlyeq m - l.$$

This time, the self-delimitation of the enumeration number of x in the B^i list comes for free, since $m - l$ is computed from m and l. All together, this amounts to

$$K^{\mathsf{B}(i+O(1))}(x) \preccurlyeq m - l + K^{\mathsf{B}(i)}(m, l). \tag{4.5}$$

Recalling that $m = K^{\mathsf{B}(i)}(x, y)$, summing (4.4) and (4.5) together yields what is to be shown.

□

The X-graph of $\mathcal{L}_x + \mathcal{L}_{y|x}$ represents the function $K^{B(i)}(x) + K^{B(i)}(y\,|\,x)$, so by Remark 3, Proposition 8 implies that $\mathcal{L}_x + \mathcal{L}_{y|x}$ is in an $O(1)$-neighbourhood of $\mathcal{L}_{x,y}$ and Proposition 9 implies that $\mathcal{L}_{x,y}$ is in an $O(\log n)$-neighbourhood of $\mathcal{L}_x + \mathcal{L}_{y|x}$. Putting this together, one has

$$\mathcal{L}_{x,y} \sim \mathcal{L}_x + \mathcal{L}_{y|x}.$$

4.3.2. Not for Description Profiles: The Antistochastic Counter-Example

In the light of the equivalence between unrelativized description and time profiles, Theorem 7, it seems that a chain rule analogous to Eq. (4.3) should also hold for description profiles. In fact,

$$\mathcal{L}_x \sim \Lambda_x \qquad \text{and} \qquad \mathcal{L}_{x,y} \sim \Lambda_{x,y},$$

so $\mathcal{L}_{x|y} \sim \Lambda_{x|y}$ holds if and only if $\Lambda_x + \Lambda_{y|x} \sim \Lambda_{x,y}$ holds. But it turns out that these relations are false in general.

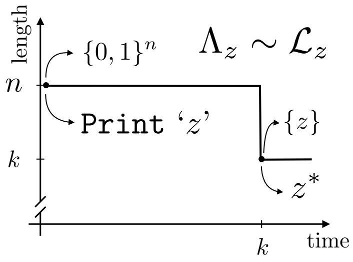


Fig. 4.2. An antistochastic string understood from the description and time profiles perspectives.

Consider the following counterexample. A string z is called *antistochastic* if its description profile contains as few elements as possible. More precisely, if $|z| = n$ and $K(z) = k$, z is ε-antistochastic if $(k-\varepsilon, n-\varepsilon) \notin \Lambda_z$. "All shapes are possible", Theorem 4, implies that there exist $O(\log n)$-antistochastic strings. Within a logarithmic precision, a profile as such is essentially generated by two points, namely, $(0,n)$ and (k,k). From the description profile perspective, these generators are witnessed by the models $\{0,1\}^n$ and $\{z\}$, respectively, while from the time perspective, those points come from the programs "`Print` z" and z^*, respectively. See Figure 4.2.

Antistochastic strings are quite strange: Every model that singles out properties of z in a more constraining way than just giving raw bits of z necessarily has complexity $\geq k$, and every program that computes z faster than $\mathsf{B}(k)$ is as long as the length of z. Even more impressive, Milovanov [62] has shown that antistochastic strings have a remarkable holographic property: If *any* $n-k$ bits of z get erased, yielding for instance

$$z' = 00*1*01**10011*00*1*111\ldots0,$$

where the "$*$" symbol represents the erased bits, then the original string can be recovered from the erased one by a logarithmic advice, *i.e.*, $K(z|z') = O(\log n)$.

Let z be such an $O(\log n)$-antistochastic string of length n and complexity k, with $n/2 < k < n$. Let $z = xy$, where the pieces x and y are chosen in such a way that each of them is insufficient to perform Milovanov's holographic reconstruction, *i.e.*, $|x| < k$ and $|y| < k$. Technically, xy does not correspond to a proper encoding of the pair $\langle x, y\rangle$ because it is not uniquely decodable, but $1^{\|x\|}0|x|xy$ is, where $\|x\|$ denotes the length of $|x|$. This discussion holds to logarithmic precision, so the prefix $1^{\|x\|}0|x|$ can be disregarded, and the profile $\mathcal{L}_{\langle x,y\rangle}$ is identified to that of $\mathcal{L}_{xy}$.

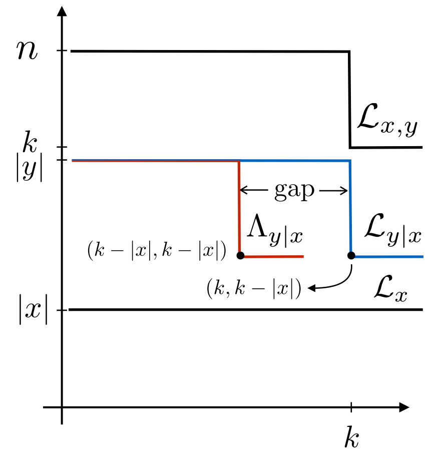


Fig. 4.3. A gap between the conditional profiles $\mathcal{L}_{y|x}$ and $\Lambda_{y|x}$ of pieces x and y of anti-stochastic strings shows that the chain rule $\mathcal{L}_{x,y} \sim \mathcal{L}_x + \mathcal{L}_{y|x}$ does not find an analogue for description profiles Λ.

Observe that x is incompressible, $K(x) \sim |x|$. Otherwise, the set $\{xw\colon w \in \{0,1\}^{|y|}\}$ would have complexity smaller than $|x|$ and with its log-cardinality of $|y|$, it would entail a two-part description smaller than $|x|+|y| = n$, contradicting the description profile. This means that, as any incompressible string, $\mathcal{L}_x$ lies just above the horizontal line of height $|x|$.

To determine $\Lambda_{y|x}$, Milovanov's property implies that

$$K(\{y\}\,|\,x) \sim K(y\,|\,x) \sim K(xy\,|\,x) \sim k - |x|,$$

and hence $(k-|x|, k-|x|) \in \Lambda_{y|x}$ (again, disregarding logarithmic precision). This point happens just after a drop since $(k-|x|-\varepsilon, |y|-\varepsilon) \notin \Lambda_{y|x}$, for some $\varepsilon = O(\log n)$. Indeed, if a program of length $k-|x|-\varepsilon$ would, from x, specify a model $S \ni y$ of log-cardinality $|y|-\varepsilon-k+|x|+\varepsilon = n-k$, then $\{xs\colon s \in S\} \ni z$ would contradict z's description profile, because it would be of unconditional complexity $k-\varepsilon$, for a two-part description of length $n-\varepsilon$.

The result of the previous section, Eq. (4.3), suffices to establish $\mathcal{L}_{y|x}$ as $\mathcal{L}_{x,y} - \mathcal{L}_x$. But for completeness, I argue it directly. It is straightforward to see that $(k, k-|x|) \in \mathcal{L}_{y|x}$: the busy running time of Milovanov's reconstruction cannot exceed the length of the program $(k-|x|)$ plus the length of the auxiliary information $(|x|)$, otherwise, it would solve too big a halting problem from too few bits. Moreover, $(k-\varepsilon, |y|-\varepsilon) \notin \mathcal{L}_{y|x}$, for some $\varepsilon = O(\log n)$ large enough. Because suppose it is: A program for y given x is then of length $|y|-\varepsilon$ and runs for $\mathsf{B}(k-\varepsilon)$ or less steps. A program of length $|x|+|y|-\varepsilon$ for z is then the "`Print` x", followed by the aforementioned program of y given x. The overall running time of such program is $\mathsf{B}(k-\varepsilon)$, hence contradicting the depth of z: Any program shorter than n that computes z must run for at least $\mathsf{B}(k)$. See Figure 4.3.

Answers typically raise more questions: Since $\Lambda_{y|x}$ and $\mathcal{L}_{y|x}$ do not coincide, what is the gap between them? As a first indicator coming from the previous example, one notices that the first coordinate of the conditional description profile is the conditional complexity (*e.g.* of value $k-|x|$) of the model, so the length of a program. However, the first coordinate of the conditional time profile is the busy running time (*e.g.* of value k) of a program, which *could be longer than its length* when auxiliary information is provided.

4.4. The Gap

On the journey towards expressing sophistication and depth of pairs, a detour is required to understand — and quantify — what separates $\Lambda_{y|x}$ from $\mathcal{L}_{y|x}$. A first step is to understand why the profiles connect in the non conditional case, and then underline what introduces a gap when relativized. To do so, I revisit the non conditional case by introducing another profile "between" $\mathcal{L}$ and Λ, which renders their link "more continuous", thus enlightening why they equate. This new profile is then relativized, allowing us to grasp what causes the gap.

4.4.1. A Man in the Middle

Levin [**56**, **75**] noticed long ago that strings with description profiles that reach $K(x)$ for large values of complexity threshold must contain mutual information with the halting problem. This is clear when such a profile is understood by its equivalent time profile, which displays programs that run for so long ($\mathsf{B}(i)$ steps!) that they can decide the halting problem for all programs shorter than i.

The following profile makes the connection with halting information even clearer. The main idea underlying its construction finds its roots in the proof by Gács [**42**] that some strings x have high $K(K(x)|x)$. The concept is further investigated by Bauwens [**10**] and named *m-sophistication*[6]. For the aware reader, here, I restrict the universal semi-measure "m" to the *a priori* probability, and I put in evidence that the quantity is a function of a significance parameter, hence defining a full-fledged profile.

To define the $\mathcal{M}_x$*-profile*, consider a dovetailed enumeration of all programs, in which each program of length j, lexicographically, is simulated during j steps of computation, for increasing values of j. Each such iteration refers to a "*j-step*". When a program of length l halts, 2^{-l} is added to a sum M initially valued at 0. Note that this process lower semi-computes Ω, so as the enumeration goes, an increasing prefix of M stabilizes to some prefix of Ω. Whenever x is produced by some program p, the current j-step is completed, and a dot is marked at the coordinate $(\theta(p),|p|)$, where $\theta(p)$ is the length of the largest prefix of Ω that has stabilized in the sum M, *i.e.*,

$$M_{[\theta]} = \Omega_{[\theta]} \qquad \text{but} \qquad M_{[\theta+1]} \neq \Omega_{[\theta+1]}.$$

I refer to $\theta(p)$, as the *time on the* Ω *clock* and the $\mathcal{M}_x$-profile is defined as the upwards and rightwards closure of the dots.

Note that $\theta(p)$ achieves the same purpose as the busy beaver renormalization $\tau(p)$, that is, it measures the running time of p in a economical representation. In fact, the two quantities are very close to one another. Indeed, the busy beaver is friends with a badger, who is also very busy.

Definition 10. *The* busy badger function $\boldsymbol{B}(i)$ *is defined as the value of the j index in the dovetailed enumeration when the first i bits of M get stabilized to $\Omega_{[i]}$. This can be written as*

$$\boldsymbol{B}(i) \overset{df}{=} \min\{j \colon M(j) = \Omega_{[i]}\beta\} \qquad \textit{(for some } \beta\textit{)},$$

where $M(j)$ denotes the value of the sum just before incrementing j to $j+1$.

The beaver and the badger can be shown to be almost as busy as one another, precisely, that

$$\mathsf{B}(i) \leq \boldsymbol{B}(i) \leq \mathsf{B}(i + K(i) + O(1)). \tag{4.6}$$

[6] Bauwens was well aware of the connection between sophistication and depth. In fact, in an earlier preprint [9], he named the concept *m-depth*. This itself is a nice wink to the "in between profiles" that is being considered here.

To see the first relation, let p be the slowest halting i-bit program, witnessing $\mathsf{B}(i)$. In the dovetailed enumeration, when p halts, the counter j has value $\mathsf{B}(i)$. Before it is supplemented by the contribution 2^{-i}, the sum M cannot have stabilized as large a prefix as $\Omega_{[i]}$, otherwise, $M + 2^{-i}$ would overshoot the value of Ω. The second relation comes from that a program hardcoded with $\Omega_{[i]}$ can execute the dovetailed enumeration and purposefully halt once i bits of Ω have stabilized. Such a program has a running time larger than $\boldsymbol{B}(i)$, but smaller than $\mathsf{B}(i + K(i) + O(1))$, because it is of length $\preccurlyeq K(\Omega_{[i]}) \preccurlyeq i + K(i)$.

By definition, the time of a program p measured on the Ω clock, is the inverse busy badger of its running time, *i.e.*, $\theta(p) = \boldsymbol{B}^{-1}(\mathrm{RT}(p))$. Recalling that $\tau(p) = \mathsf{B}^{-1}(\mathrm{RT}(p))$, inverting the relation (4.6) yields

$$\tau(p) \lesssim \theta(p) \le \tau(p). \tag{4.7}$$

This connection between τ and θ establishes the first relation of the following statement, which is in its whole a corollary of the upcoming Propositions 15 and 16. It states that the $\mathcal{M}$-profile can indeed be considered as *a man in the middle* between the $\mathcal{L}$-profile and the Λ-profile.

Corollary 11. *For all x,*

$$\mathcal{L}_x \sim \mathcal{M}_x \sim \Lambda_x. \tag{4.8}$$

A Computable Shape

The following proposition states that the shape of the $\mathcal{M}_x$-profile can be precisely computed from x^* and its time on the Ω clock.

Proposition 12. *For all x,*

$$\begin{tikzcd}[row sep=tiny]
x^* \arrow[dr] & & \\
& \boxed{O(1)} \arrow[r] & \mathcal{G}(\mathcal{M}_x) \\
\theta(x^*) \arrow[ur] & & .
\end{tikzcd}$$

Proof. From x^* and θ, one can compute $\Omega_{[\theta]}$ by executing the dovetailed enumeration until x^* halts in the enumeration. Thanks to θ, one then knows what is the precise prefix of M that has been stabilized, $\Omega_{[\theta]}$. With this at hand, one then starts again the dovetailed enumeration, this time, marking a dot at the coordinate (i', l) when a program of length l has computed x in time i' on the Ω clock. One can then return any finite representation of $\mathcal{M}_x$, for instance, the minimal one $\mathcal{G}(\mathcal{M}_x)$. □

The previous proposition makes more specific a result by Vereshchagin and Vitányi [75, §7], which informally states that from x, $K(x)$ and the complexity of a near minimal sufficient statistics, a curve λ' can be computed, whose closure is logarithmically close to Λ_x. By the above, this λ' can be taken to be $\partial\mathcal{M}_x$.

4.4.2. Relativizing the $\mathcal{M}$-Profile

Halting information, too, can be relativized to some auxiliary information z, since it is in general uncomputable to determine whether a program p yields a halting $\mathcal{U}(p,z)$. This *relative halting information* can again take the form of a halting probability,

$$\Omega^z \overset{\text{df}}{=} \sum_{p\,:\,\mathcal{U}(p,z)\searrow} 2^{-|p|}.$$

By a similar argument as in the unconditional case (*cf.* Section 4.2.1), $\Omega^z_{[i]}$ solves the halting problem relative to z, for all programs of length $\leq i$. Thus, $i \preccurlyeq K(\Omega^z_{[i]}|z) \preccurlyeq i + K(i)$.

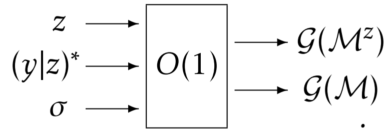
Two ways!

It turns out that the $\mathcal{M}$-profile can be relativized in two natural ways. To define these conditional profiles, think of a dove with two tails. One dovetailed enumeration runs all programs of length $\leq j$, lexicographically, for j steps of computation on the reference computer $\mathcal{U}$, *supplemented by auxiliary information* z. If $\mathcal{U}(p,z) \searrow$ for some program p, then $2^{-|p|}$ is added to a sum M^z. Before incrementing the j counter, the other tail is visited, running the same programs, also for j steps of computation on $\mathcal{U}$, but *without auxiliary information*. If $\mathcal{U}(q) \searrow$ for some program q, then $2^{-|q|}$ is added to a different sum M. I shall refer to $M^z(j)$ and $M(j)$ as the value taken by the sums just before incrementing the counter to $j+1$. The idea is to have two clocks to measure time: one follows the stabilization of a prefix of Ω^z by M^z and the the other, of Ω by M.

Whenever the first enumeration finds a p such that $\mathcal{U}(p,z) = y$, the j^{th} step is completed and a red dot is marked at the coordinate $(\theta^z(p), |p|)$, where $\theta^z(p)$ is the length of the largest prefix of Ω^z that have stabilized in $M^z(j)$. I shall call $\theta^z(p)$ the *time on the* Ω^z *clock*. Additionally, a blue dot is marked at the coordinate $(\theta(p), |p|)$, where $\theta(p)$ is as before, the time on the Ω clock. Define $\mathcal{M}^z_{y|z}$ and $\mathcal{M}_{y|z}$ as the upwards and rightwards closure of the red and blue dots, respectively.

Remark 13. $\mathcal{M}^{\varepsilon}_{y|\varepsilon} = \mathcal{M}_{y|\varepsilon} = \mathcal{M}_y$.

Proposition 14. *Let σ be the time* either *on the Ω or on the Ω^z clock when $(y|z)^*$ halts, then*



Proof. The proof is analogous to that of Proposition 12. □

Each with his Own Mate

The following two propositions state that each version of the relative $\mathcal{M}$-profiles follows its own other relative profile:

$$\mathcal{M}^z_{y|z} \sim \Lambda_{y|z} \qquad \text{while} \qquad \mathcal{M}_{y|z} \sim \mathcal{L}_{y|z}\,.$$

Proposition 15. *For all y and z,*

$$\mathcal{L}_{y|z} \subseteq \mathcal{M}_{y|z} \qquad \textit{and} \qquad \mathcal{M}_{y|z} \subseteq O(\log n)\textit{-neighbourhood of } \mathcal{L}_{y|z}\,.$$

This implies that $\mathcal{L}_{y|z} \sim \mathcal{M}_{y|z}$.

Proof. A program p is the fastest witness of a point in $\mathcal{G}(\mathcal{L}_{y|z})$ if and only if it is the fastest witness of a point in $\mathcal{G}(\mathcal{M}_{y|z})$, both points being horizontally aligned at height $|p|$. By taking such a fastest witness for $(\tau(p), |p|) \in \mathcal{G}(\mathcal{L}_{y|z})$ and $(\theta(p), |p|) \in \mathcal{G}(\mathcal{M}_{y|z})$, the conclusion follows from Equation (4.7) and remark 3 (i). □

Proposition 16. *For all y and z,*

$$\begin{aligned} \Lambda_{y|z} &\subseteq O(1)\textit{-neighbourhood of } \mathcal{M}^z_{y|z} \qquad \textit{and} \\ \mathcal{M}^z_{y|z} &\subseteq O(\log n)\textit{-neighbourhood of } \Lambda_{y|z}\,. \end{aligned}$$

This means that $\Lambda_{y|z} \sim \mathcal{M}^z_{y|z}$.

Proof. $\Lambda_{y|z} \subseteq O(1)$*-neighbourhood of* $\mathcal{M}^z_{y|z}$.
Let $S \ni y$ be a model witnessing $(i, \lambda) \in \mathcal{G}(\Lambda_{y|z})$. It induces a program that computes y from z via its two-part description, of length λ. But what is its time on the Ω^z clock? Being i-bit long, the first part runs for at most time $i + O(1)$ on the Ω^z clock, which is the most conservative bound for an i-bit program running with auxiliary information

z. And *the second part of a two-part description is fast*: It takes $O(|S|) \leq O(|\{0,1\}^{|y|}|)$ steps, which is negligible compared to the time bound of the first part, so it can be absorbed by increasing the time on the Ω^z clock to $i + O(1)$.

$\mathcal{M}^z_{y|z} \subseteq O(\log n)$-neighbourhood of $\Lambda_{y|z}$.

Let $(i,\ell) \in \mathcal{G}(\mathcal{M}^z_{y|z})$ be witnessed by a program p computing y given z, of length ℓ and time i on the Ω^z clock. Programs can be grouped together on the basis of their length and their time on the Ω^z clock. Hence, for arbitrary l, define

$$\begin{aligned}
\tilde{A}_{i,l} &= \{r \colon |r| = l \text{ and } \theta^z(r) \geq i\}, \\
\bar{A}_{i,l} &= \{r \colon |r| = l \text{ and } \theta^z(r) = i\} \text{ and} \\
A_{i,l} &= \{\mathcal{U}(r) \colon |r| = l \text{ and } \theta^z(r) = i\}.
\end{aligned}$$

Notice that p is an element of the first two sets and that y is an element of the $A_{i,\ell}$. I shall show that $A_{i,\ell}$ is a model with

$$K(A_{i,\ell}|z) \lesssim i \qquad \text{and} \qquad K(A_{i,\ell}|z) + |A_{i,\ell}| \lesssim \ell\,.$$

First, observe that given z, $A_{i,\ell}$ can be computed from $\bar{A}_{i,\ell}$, which can be computed from $\Omega^z_{[i]}$ and ℓ, so

$$K(A_{i,\ell}|z) \preccurlyeq K(\Omega^z_{[i]}, \ell|z) \lesssim i\,.$$

Second, the log-cardinality of $A_{i,\ell}$ needs to be bounded, and because it contains fewer elements than $\tilde{A}_{i,l}$, bounding the latter suffices. Define $a_{il} \equiv |\tilde{A}_{i,l}|$. For a fixed i, the discrete application $l \mapsto a_{il}$ is lower semi-computable from z and $\Omega^z_{[i]}$. Moreover,

$$\sum_l a_{il} 2^{-l} \leq 2^{-i}\,,$$

otherwise, too large of an algorithmic mass of programs would remain to halt — contradicting the i-th bit of Ω. This means that $a_{il}2^{-l+i}$ is a lower semi-computable semi measure, relative to z and $\Omega^z_{[i]}$, so by the coding theorem[7],

$$a_{il} \leq 2^{l-i-K(l|z,\Omega^z_{[i]},K(\Omega^z_{[i]}|z))+O(1)}\,.$$

[7] The coding theorem [59] states that every discrete application $j \mapsto \mu(j)$ that is (i) lower semi-computable from auxiliary information z and (ii) a semi-measure, *i.e.*, $\sum_j \mu(j) \leq 1$, has $\mu(j) \leq 2^{-K(j|z)+O(1)}$.

Therefore, $\log|A_{i,\ell}| \leq \log a_{i\ell} \preccurlyeq \ell - i - K(\ell\,|\,z, \Omega^z_{[i]}, K(\Omega^z_{[i]}|z))$. So

$$\begin{aligned} K(A_{i,\ell}) + \log|A_{i,\ell}| &\preccurlyeq K(\Omega^z_{[i]}, \ell\,|\,z) + \ell - i - K(\ell\,|\,z, \Omega^z_{[i]}, K(\Omega^z_{[i]}|z)) \\ &\asymp K(\Omega^z_{[i]}\,|\,z) + \ell - i \\ &\preccurlyeq \ell + K(i) \qquad (4.9) \\ &\lesssim \ell. \end{aligned}$$

□

Corollary 11 thus follows from the last two propositions and from Remark 13. This corollary is the equivalence between depth and sophistication. Minimal sufficient statistics induce a two part-code in a way that forces triviality, and hence fast computation, of the second part. This then distillates all the slow computation (*i.e.*, the deep structures in Bennett's sense) into the model (the sophisticated structures in Kolmogorov's sense). In an algorithmic information theoretic sense, those deep and sophisticated structures essentially made of initial segments of Ω; they are full of halting information.

4.4.3. Losing Synchronicity

Light can now be shed on the difference between the conditional profiles $\mathcal{L}_{y|z}$ and $\Lambda_{y|z}$, through their equivalent representations in terms of conditional $\mathcal{M}$-profiles. The difference between $\mathcal{M}^z_{y|z}$ and $\mathcal{M}_{y|z}$ is only a horizontal distortion, since generators come in horizontally aligned pairs as they are witnessed by the same program p whose length establishes the second coordinate. The distortion reflects that of the clocks Ω^z and Ω, with respect to which the times $\theta^z(p,z)$ and $\theta(p,z)$ determine the first coordinate.

As a first step to better characterize the difference in the flow of the clocks, the program p witnessing the aforementioned aligned generators can be abstracted, and rely only on the times θ^z and θ showed by the clocks.

Definition 17. *Define the* relativized busy badger, $\boldsymbol{B}_z(i)$, *as the value of the j index when i bits of Ω^z have stabilized, namely,*

$$\boldsymbol{B}_z(i) \overset{df}{=} \min\{j\colon M^z(j) = \Omega^z_{[i]}\beta\}.$$

Observe that

$$\boldsymbol{B}_z(\theta^z) \leq \mathrm{RT}(p,z) < \boldsymbol{B}_z(\theta^z+1) \implies \boldsymbol{B}^{-1}\boldsymbol{B}_z(\theta^z) \leq \theta \leq \boldsymbol{B}^{-1}\boldsymbol{B}_z(\theta^z+1),$$

so the connection from the Ω^z to the Ω clock is $\boldsymbol{B}^{-1}\boldsymbol{B}_z(\cdot)$. In what follows, this connection is reframed in terms of wether — and if so *how* — z has information about the halting problem.

Relation with Halting Knowledge

First, I exemplify this connection. Suppose $z = \Omega_{[a]}$, and $\theta^z = b$, what is the corresponding time θ on the Ω clock? If b bits of Ω^z have stabilized, it means that no more programs shorter than b bits in length will ever lead to a halting $\mathcal{U}(\cdot,z)$ computation, in particular, the program described in the following paragraph.

With the help of z and some extra hardcoded bits of Ω, assemble $\Omega_{[a+b-O(\log b)]}$, and execute the dovetailed enumeration of programs, run without z, until the sum exceeds $\Omega_{[a+b-O(\log b)]}$. This particular computation takes a time $a+b-O(\log b)$ on the Ω clock, so $\theta \gtrsim a+b$. By incompressibility of Ω, in fact $\theta \sim a+b$ holds. This example puts in evidence that the gap between θ^z and θ depends upon z's knowledge about the halting problem. More precisely, the distortion in the flow of the clocks turns out to be a property of *the manner* in which z has such knowledge.

In the spirit of the above example, the following definition quantifies how close to Ω one can get from z and i bits of advice.

Definition 18. *The* reach curve *of z is defined as*

$$Reach_z(i) \overset{df}{=} \max\{s\colon K(\Omega_{[s]}\beta\,|\,z) \leq i \textit{ and } \Omega_{[s]}\beta < \Omega\}.$$

This definition is reminiscent of monotone complexity, where the finite string β is a tool for an overall $\Omega_{[s]}\beta$ possibly simpler than the raw $\Omega_{[s]}$, generally for length reasons. The terminology has a twofold interpretation. $\mathrm{Reach}_z(i)$ measures how close to Ω can be reached, which is directly connected to how large a number (or running time) can be reached. If z is independent from the halting problem, its reach curve follows the identity line within logarithmic resolution: i bits of program grants $\sim i$ bits of prefix of

Ω. However, if z contains pieces of information about Ω its reach curve will display the benefits of that knowledge by moving above of the identity line.

The following proposition pinpoints *what* information is the most helpful for z to reach as large a prefix of Ω as possible. In other words, what should the i bits of advice be made of? The answer is the initial bits of Ω^z. Hence, if z has holes in its halting knowledge, then Ω^z fills them.

Proposition 19. *Let* $\mathrm{Reach}_z(i) = r$ *be witnessed by the program* p *such that* $\mathcal{U}(p,z) = \Omega_{[r]}\beta < \Omega$ *and* $|p| \leq i$. *Then* p*'s algorithmic information is essentially that of* $\Omega^z_{[i]}$, *since*

$$K(\Omega_{[r]}\beta \,|\, z, \Omega^z_{[i]}) = O(1).$$

Proof. From z and $\Omega^z_{[i]}$, one can compute the list $\mathcal{U}(q,z)$, for all halting programs q of length $\leq i$ (the non-halting programs are discarded). Each such q can be transformed in a program $O(1)$ longer that I shall call the the $\mathcal{U}(q,z)$-dovetail. This consists of the dovetailed enumeration of all programs, run without z, until the sum M exceeds the value of $\mathcal{U}(q,z)$ previously computed. The halting status of each $\mathcal{U}(q,z)$-dovetail can be obtained from z and $\Omega^z_{[i+O(1)]}$. The latter is non-constructively acquired by the $O(1)$ advice. The largest $\mathcal{U}(q,z)$ leading to a halting $\mathcal{U}(q,z)$-dovetail is then outputted. □

The next proposition states that the reach curve $\mathrm{Reach}_z(i)$ expresses equivalently the connection $\boldsymbol{B}^{-1}\boldsymbol{B}_z(i)$ between clocks: they are logarithmically close to one another. Both relations are non-decreasing, so their upwards and *leftwards* closure define the respective profiles $\mathcal{R}_z$ and $\mathcal{B}^{-1}\mathcal{B}_z$. Since the profiles can go beyond the length of z, the upcoming "$\sim$" relation refers to $O(\log i)$, where i is the first coordinate of the profiles' points.

Proposition 20. *For all* z,

$$\mathcal{R}_z \sim \mathcal{B}^{-1}\mathcal{B}_z.$$

Proof. It suffices to show that $\mathrm{Reach}_z(i') \geq \boldsymbol{B}^{-1}\boldsymbol{B}_z(i)$, for $i' \lesssim i$ and vice versa. Let $\boldsymbol{B}^{-1}\boldsymbol{B}_z(i) = r$, so when the double dovetailed enumeration is performed, when i bits of Ω^z stabilize, r bits of Ω are stabilized. A program of length $\lesssim i$, with knowledge

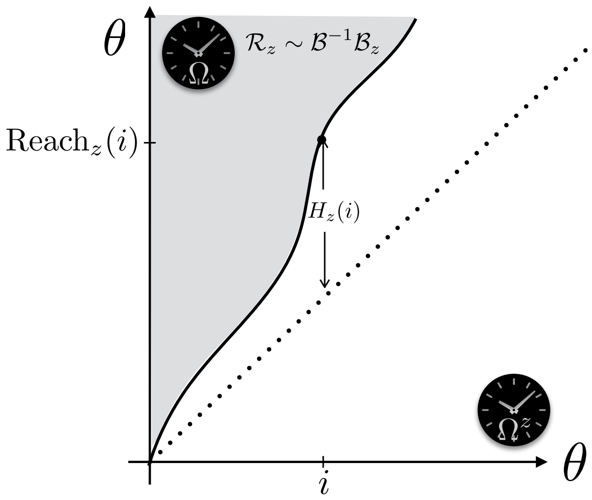


Fig. 4.4. The connection between the Ω and the Ω^z clock is given by the reach curve of z.

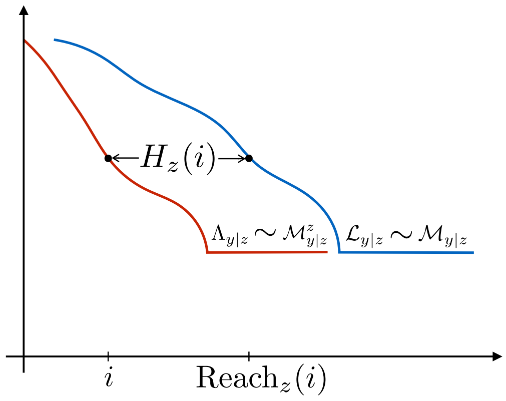


Fig. 4.5. The conditional profiles and the gap between them.

of $\Omega^z_{[i]}$, can then compute $\Omega_{[r]}\beta$ by also running the two dovetailed enumerations, and when $\Omega^z_{[i]}$ is stabilized in one enumeration, outputs the sum $M = \Omega_{[r]}\beta$ of the other[8].

Now, I show that $\boldsymbol{B}^{-1}\boldsymbol{B}_z(i') \geq \mathrm{Reach}_z(i)$, for $i' \lessdot i$. Let $\mathrm{Reach}_z(i) = r$, so $\Omega_{[r]}\beta = \mathcal{U}(p, z)$ for some p of length $\leq i$. This program can be transformed into the $\Omega_{[r]}\beta$-dovetailing, of length $i' = i + O(1)$. Therefore $\boldsymbol{B}_z(i')$ is no smaller than the running time of the $\Omega_{[r]}\beta$-dovetailing, which is long enough to stabilize r bits on the Ω clock, so $\boldsymbol{B}_z(i') \geq \boldsymbol{B}(r)$. □

Naming the Gap

Of interest is the quantity

$$H_z(i) \overset{\text{df}}{=} \text{Reach}_z(i) - i\,,$$

which measures the time *difference* between the Ω and the Ω^z clocks, hence, the gap between the relative profiles. See Figures 4.4 and 4.5. Being an affine transformation of $\text{Reach}_z(i)$, it encodes the same information.

H_z may be called the *halting materialization distribution* because of the following observations. For small values (logarithmic in the length of z), the halting materialization distribution coincides with the reach curve,

$$H_z(O(\log n)) \sim \text{Reach}_z(O(\log n))\,,$$

and represents the largest prefix of Ω that can be computed from a logarithmic advice (such halting information materializes easily). This value is an important characteristic of strings with any sort of interesting profiles. In fact, a string z that displays a drop at value d in his time profile $\mathcal{L}_z$, will have $H_z(O(\log n)) \gtrsim d$. This is because such a drop witnesses that the fastest program of a certain length ℓ, that computes z, runs for $\mathsf{B}(d)$ steps of computation, long enough to stabilize almost d bits of Ω. Therefore, with z at hand, an $O(\log n)$ advice to reach close to Ω is simply *"ℓ"*. It serves as a promise of finding an ℓ-bit long program that computes z. In the process of finding it, the sum M of the dovetailed enumeration will stabilize $\sim d$ bits of Ω.

And at the other end of the spectrum, $\lim_{i\to\infty} H_z(i) = I(z\colon \Omega)$. In fact,

$$\begin{aligned}
H_z(i) &= \max\{s - i\colon K(\Omega_{[s]}\beta\,|\,z) \le i\} \\
&\sim \max\{s - K(\Omega_{[s]}\beta\,|\,z)\colon K(\Omega_{[s]}\beta\,|\,z) \le i\} \\
&\sim \max\{K(\Omega_{[s]}) - K(\Omega_{[s]}\,|\,z)\colon K(\Omega_{[s]}\beta\,|\,z) \le i\} \\
&\sim \max\{I(\Omega_{[s]}\colon z)\colon K(\Omega_{[s]}\beta\,|\,z) \le i\}\,.
\end{aligned}$$

As i grows, s grows at the same pace or faster, so in the limit $i \to \infty$, s goes also to ∞. In between small and large values, the shape of H_z informs us of how hard it is to materialize

[8] Can it be shown that the monotone complexity of $\Omega_{[i]}$ is smaller than $i + O(1)$, *i.e.*, $\forall i \exists \gamma K(\Omega_{[i]}\gamma) \preccurlyeq i$? If so the 2 profiles would be $O(1)$ close.

the halting knowledge of z. For instance, ζ made of bits number 501 to 2000 of Ω is useless to solve *any* halting problem... until a clever 500-bit advice is provided. In such a case, the halting information of ζ is only *materialized* after $i = 500$, and H_ζ is indeed a step function, with the step at that value.

If the antistochastic strings looked like the strangest of all in the view of their Λ and their $\mathcal{L}$ profiles, still, they display a relatively straightforward halting materialization distribution: It is constant at the value corresponding to the drop of the $\mathcal{L}$ profile, which is at value of their complexity. More elaborate halting materialization profiles are possible and in fact, the following proposition shows that all shapes are possible.

Proposition 21. *For any non-decreasing function $h(i)$ that eventually remains constant, there exist a string γ whose halting materialization distribution $H_\gamma(i)$ is $O(K(h))$ close to $h(i)$, where $K(h)$ is the complexity of the function h, which is defined as* $\min_p\{|p|\colon p \text{ computes } h\}$.

Proof. This proof is about playing a game with the bits of Ω, in which one basically encodes the graph of $h(i)$ into γ, as to which bits of Ω are given. Let a_0 be the first integer mapped to a non-null value, $b_0 = h(a_0)$. Let $a_1, a_2, \dots, a_m$ all the values at which h increases, and $b_1, b_2, \dots, b_m$ the corresponding amounts by which h increases. Define

$$\gamma \equiv \Omega_{a_0}\dots\Omega_{a_0+b_0}\Omega_{a_0+b_0+a_1}\dots\Omega_{a_0+b_0+a_1+b_1}\Omega_{a_0+b_0+a_1+b_1+a_2}\dots\Omega_{\sum_0^m a_i \sum_0^m b_i}\,,$$

where Ω_c stands for the c-th bit of Ω. From γ and $\asymp i + K(h,i)$ bits of advice, a prefix of Ω is obtained by "patching its holes" with a string of length i defined as

$$\delta = \Omega_1\dots\Omega_{a_0-1}\Omega_{a_0+b_0+1}\dots\Omega_{a_0+b_0+a_1-1}\dots\Omega_{h(i)+i-1}\Omega_{h(i)+i}\,.$$

The extra $K(h,i)$ bits are required for delimitation purposes: Not only self-delimitation of δ, but mostly to unravel the bits of γ and the bits of δ in order to assemble $\Omega_{[i+h(i)]}$. This particular choice of advice shows that

$$\mathrm{Reach}_\gamma(i + K(h,i) + O(1)) \geq i + h(i)\,.$$

A program of such a length could not compute a larger prefix, since it would contradict the incompressibility of Ω. □

Let me return to where we started. The conditional profiles $\Lambda_{y|z}$ and $\mathcal{L}_{y|z}$ do not correspond. They have been shown to be equivalently represented by the profiles $\mathcal{M}^{z}_{y|z}$ and $\mathcal{M}_{y|z}$, respectively, whose difference is a horizontal distortion quantified by $H_z(i)$. This distortion measures the difference of flow between the Ω^z and the Ω clocks, which is related to the difficulty of z to materialize its halting information in terms of a prefix of Ω.

4.5. Depth and Sophistication of Pairs

As mentioned in Section 4.3, a cornerstone of the algorithmic theory of information is the chain rule, eq. (4.1), which relates the complexity of a pair to that of a single string and a conditional homologue. Logical depth and sophistication arose from an effort to measure the meaningful information in a string, and not just its randomness. In the light of the previous results, depth and sophistication of pairs can now be expressed in terms of their single string and conditional versions.

For tidier expressions characterizing depth and sophistication for pairs, one should free the concepts from their significance parameters, keeping only the essence of what they capture. This is achieved when the significance parameters are taken as small as possible.

4.5.1. Depth$_0$

For the busy beaver depth, the natural candidate of a parameter-free version is $\mathrm{Depth}^{\mathsf{B}}_0(\cdot)$. It amounts to the busy running time of the (fastest) shortest program. The significance parameter of the busy beaver depth can meaningfully be taken to 0, because time profiles are not naturally bumpy: Even the smallest drop of one unit deep in the $\mathcal{L}_x$ profile of some string x is very significant. Such a drop grasps that x contains a lot of mutual information with a prefix of Ω, simply through the running time of its shortest program.

However, such a micro drop as the last drop of the profile is problematic in the task of formulating a relation between $\mathrm{Depth}^{\mathsf{B}}_0(x,y)$, $\mathrm{Depth}^{\mathsf{B}}_0(x)$ and $\mathrm{Depth}^{\mathsf{B}}_0(y\,|\,x)$, since the main tool at hand is the relation (4.3), $\mathcal{L}_{x,y} \sim \mathcal{L}_x + \mathcal{L}_{y|x}$, which incorporates errors of logarithmic order on the Y axis. Recall that the depth profile $\mathcal{D}$, Eq. (4.3), is a downwards translation of the time profile $\mathcal{L}$, with $\mathrm{Depth}_c^{\mathsf{B}}$ being represented as the Y*-graph* of $\mathcal{D}$. Consequently, the errors of logarithmic order transpose on the axis of the depth's significance parameter. To keep the discussion grounded in the ideas, I will avoid the conundrum by imposing an extra constraint on the considered profiles. The strings x and y are said to have $\mathcal{L}$-profiles with a *sharp finish* if all their time profiles (*e.g.*, $\mathcal{L}_{x,y}$, $\mathcal{L}_{x|y}$, $\ldots$) display a last drop that is greater than some $\varepsilon = O(\log n)$. More precisely, the parameter ε is chosen greater than the sum of the error terms in Propositions 8 and 9. This ensures that the latest drop of $\mathcal{L}_{x,y}$ is aligned (up to $O(1)$ resolution) with either the latest drop of $\mathcal{L}_x$ or with the latest drop of $\mathcal{L}_{y|x}$. The X coordinate at which the latest drops happen marks the $\mathrm{Depth}_0^{\mathsf{B}}$. Therefore, if x and y have profiles with a sharp finish,

$$\mathrm{Depth}_0^{\mathsf{B}}(x,y) \asymp \max\{\mathrm{Depth}_0^{\mathsf{B}}(x), \mathrm{Depth}_0^{\mathsf{B}}(y\,|\,x)\}. \tag{4.10}$$

This relation means that the running time (in busy beaver units) of the shortest program that produces the pair x, y is close to either that of x^* or that of $(y|x)^*$. Since the relation can instead be developed on y and $x|y$, if x or y is deep, so is the pair. However, the reciprocal does not hold. When x and y are pieces of an antistochastic string, each of them is individually shallow but deep relative to one another, yielding a deep pair.

Finally the deterministic *slow growth law* [16] can be retrieved from Equation (4.10). In fact, let y be a computable processing of x. If x is shallow, but y is deep, then the relation implies that y is deep relative to x: it cannot have been computed by a short *and* fast program.

4.5.2. A Parameter-free Sophistication?

Exhibiting a parameter-free notion of sophistication is a more sophisticated task ;-). In an aphorism, sophistication is the complexity of the minimal sufficient statistic, but then, what is the precise criterion for a statistic to be sufficient? A sufficient statistic is often (*e.g.*, [75, §2] [77, §5] [43]) defined to be an $S \ni x$ that satisfies

$$K(S) + \log|S| = K(x) + O(1). \tag{4.11}$$

However, the nature of two-part descriptions generally makes this relation too difficult to satisfy.

Before I elaborate more on this, I must mention that Antunes and Fortnow [4] approached the problem of liberating sophistication from its parameter by including it in the minimization. *Coarse sophistication* is thus defined as

$$\mathrm{cSoph}(x) = \min_c \{\mathrm{Soph}_c(x) + c\}.$$

This definition suffers from the problem that it does not do justice to the most sophisticated strings of a fixed length n. Indeed those have an antistochastic-like profile, with a drop (of height $\delta = n - K(x)$) as late as possible (at $K(x)$). A late drop as such forces $K(x)$ to be close to n, thereby shrinking the height δ of the drop. Consider a string x as such with δ small, but still in $\Omega(n)$. Its sophistication is large: $\mathrm{Soph}_c(x) = K(x)$, for $c \leq \delta$, as witnessed by its only minimal sufficient statistic $\{x\}$. However, its coarse sophistication collapses to δ, as witnessed by $\{0,1\}^n$.

I come back to the perhaps too strict constraints of the criterion of Eq. (4.11). As mentioned in the preliminaries, the shortest one-part description for x, this is x^*, in itself carries more algorithmic information than x alone: It carries its own length $K(x)$,

$$\begin{tikzcd} & & x \\ x^* \arrow[r] & \boxed{O(1)} \arrow[ur] \arrow[dr] & \\ & & K(x)\,. \end{tikzcd}$$

For the same self-delimitation reason, a two-part description $D(S^*, i_S^x) = \alpha S^* i_S^x$ carries in itself two implicit lengths: those of each part. Thereby,

$$\begin{tikzcd} & & x \\ S^* \arrow[dr] & & \\ & \boxed{O(1)} \arrow[uur] \arrow[r] \arrow[ddr] & K(S) \\ i_S^x \arrow[ur] & & \\ & & \log|S|\ , \end{tikzcd}$$

so $K(S) + \log|S| \succcurlyeq K(x, K(S), \log|S|) \asymp K(x) + K(K(S), \log|S| \,|\, x, K(x))$.

For $K(S) + \log|S|$ in the vicinity of $K(x)$, the extra complexity brought by the last term is essentially that of a delimiter, $K(S)$, that breaks the number $K(x)$ in two pieces. Arguments can be made that by increasing the value of that delimiter, it will eventually be of small complexity, given $K(x)$. But can this "small" be qualified to be $O(1)$? No, since in general the exact value of this complexity cannot be set uniformly for all x, except, obviously, when the delimiter reaches the end of the spectrum, $K(S) \asymp K(x)$, with $S = \{x\}$. Therefore, the tail of the Λ-profile is not smooth, since unlike with the $\mathcal{L}$-profile, small deeper drops may meaninglessly occur. Indeed, these may simply be an artifact of a model $S \ni x$ with larger $K(S)$, but with smaller $K(K(S) \,|\, x, K(x))$.

Hence, a parameter-free notion of sophistication should accommodate the fact that $K(S)$ is in general completely independent from the algorithmic information of $K(x)$. For instance, in the proof of Proposition 16, where a model of x was built from a program that computes x, the length of the two-part description was large enough to encompass the complexity of the delimiter between each part of the description. In fact, this can be seen from Equation (4.9), which reduces to

$$K(A) + \log|A| \preccurlyeq K(x) + K(i) \text{ with } i \sim K(A),$$

if $z = \epsilon$ (no auxiliary information) and $\ell = K(x)$ (build the shortest two-part description from the shortest program). Therefore as a candidate for a parameter-free sophistication, one could take

$$\min_{S \ni x}\{K(S) \colon K(S) + \log|S| \leq K(x) + K(K(S)) + O(1)\},$$

which is guaranteed to be witnessed early enough by the two-part description built in the proof of 16, for appropriate choice of Proposition $O(1)$. However, if we are to rely on *the proof of the equivalence* between Λ and $\mathcal{M}$ to define sophistication without parameters, we might as well rely on *the equivalence itself*. Like $\mathcal{L}$, and unlike Λ, $\mathcal{M}$ has a smooth, constant tail of profile, which enables a meaningful definition at 0 bits of significance.

I then define the *parameter-free sophistication*, and its conditional homologue, as

$$\begin{aligned}\mathrm{Soph}(x) &\overset{\mathrm{df}}{=} \min\{i \colon (i, K(x)) \in \mathcal{M}_x\}\\ \mathrm{Soph}(y\,|\,z) &\overset{\mathrm{df}}{=} \min\{i \colon (i, K(y\,|\,z)) \in \mathcal{M}^z_{y|z}\}.\end{aligned}$$

The unconditional version coincides with Bauwens's [**10**] m-sophistication[9] k_0, and within logarithmic precision, with $\mathrm{Depth}^{\mathsf{B}}_0$. The conditional version, however, follows $\mathcal{M}^z_{y|z} \sim \Lambda_{y|z}$ instead of $\mathcal{L}_{y|z}$.

With these definitions at hand, the results of Section 4.4 imply that if x and y have $\mathcal{M}$-profiles with sharp finish,

$$\begin{aligned}\mathrm{Soph}(x,y) &\asymp \max\{\mathrm{Soph}(x),\ \mathrm{Reach}_x(\mathrm{Soph}(y\,|\,x))\} \qquad (4.12)\\ &= \max\{\mathrm{Soph}(x),\ \mathrm{Soph}(y\,|\,x) + H_x(\mathrm{Soph}(y\,|\,x))\}.\end{aligned}$$

[9] With the *a priori* probability as a universal semi-measure.

Recall the example of Section 4.3.2, showcasing an antistochastic string $z = xy$. The gap between the conditional profiles illustrated in Figure 4.3 can now be understood in terms of the halting materialization distribution $H_x(i)$, evaluated at $i = k - |x|$, which consistently amounts to $|x|$. Indeed, from x and $\sim k - |x|$ bits of advice (taken from y), Milovanov's reconstruction of z can be performed. By the shape of $\mathcal{L}_z$, such a short program must run for at least a busy running time of k, which is long enough to stabilize $\gtrsim k$ bits of Ω. Therefore $\mathrm{Reach}_x(k - |x|) \sim k$ so $H_x(k - |x|) \sim |x|$.

Finally, the parameter-free depth and sophistication correspond to the first coordinate of the "bottom left corners" of each profile displayed on Figure 4.3. Recalling that $\mathcal{L}_x \sim \Lambda_x$ and $\mathcal{L}_{x,y} \sim \Lambda_{x,y}$, one finds

$$\begin{aligned} &\mathrm{Depth}_0^{\mathsf{B}}(x,y) \sim k\,, \qquad &&\mathrm{Depth}_0^{\mathsf{B}}(x) \sim 0\,, \qquad &&\mathrm{Depth}_0^{\mathsf{B}}(y\,|\,x) \sim k\,, \\ &\mathrm{Soph}(x,y) \sim k\,, &&\mathrm{Soph}(x) \sim 0 \quad \text{and} &&\mathrm{Soph}(y\,|\,x) \sim k - |x|\,. \end{aligned}$$

And the established relations (4.10) and (4.12) are easily verified.

4.6. Conclusions

The goal has been reached. Thanks to the time profile chain rule of §4.3, the busy beaver depth of a pair $\mathrm{Depth}_0^{\mathsf{B}}(x,y)$ can be expressed in terms of $\mathrm{Depth}_0^{\mathsf{B}}(x)$ and $\mathrm{Depth}_0^{\mathsf{B}}(y\,|\,x)$, simply as their maximum. Had the equivalence of depth and sophistication been carried over by the relative case, it would have been straightforward to formulate a sophistication analogue. The nature of the gap between relative depth and relative sophistication was enlightened in the detour of §4.4. Best journeys have detours; it turns out that this gap reveals more subtle structures in a string x than those expressed by the Kolmogorov structure function, equivalently represented by Λ_x or $\mathcal{L}_x$. In fact, the halting materialization distribution H_x expresses the ability — or the difficulty — for x to solve the halting problem from advices of increasing size.

The antistochastic string z — and pieces x and y of it — served as an anchor throughout the paper. Although x has the same Kolmogorov structure function as any incompressible string, its halting materialization distribution H_x is very different from that of typical strings: it knows about the halting problem — and in a somewhat peculiar way. With not enough bits of advice, x is useless to solve *any* halting problem. However, with a large enough advice, its irreducible halting information, *i.e.*, all of its algorithmic information, becomes useful. For more on antistochastic strings, see §4.7.

The Irrelevant Oracle Problem [63].

The gist of that problem can be formulated as follows. From a pair of strings (x,y), c bits of *common information can be extracted*, for a threshold t, if there exists γ such that

$$K(\gamma|x) < t, \qquad K(\gamma|y) < t \qquad \text{and} \qquad K(\gamma) \geq c.$$

Assume that $I(\langle x,y\rangle : z) \sim 0$. Can this z (an apparently irrelevant oracle) help to extract common information between x and y, *e.g.*, by altering the values t and c in the relativized case? Muchnik and Romashchenko [63] have provided a negative answer when x and y are *stochastic* strings, that is, their Λ profile contains as many points as possible. But the general case is still open. Can the halting materialization distribution find an application to the problem?

Depth from Expectation.

The logical depth of x is defined as the running time of its most probable programs, namely, the shorter ones. This allows us to ignore the fast but long programs, such as the "`Print` x" program. But if such an origin is anyways algorithmically improbable, why not defining the logical depth as the *expected* running time of the computational origines of x; with expectation taken over the algorithmic probability? Something like

$$\sum_{p:\mathcal{U}(p)=x} 2^{-|p|}\mathrm{RT}(p)\ ?$$

It is a nice try, but it makes no sense since this sum diverges for all x. Indeed, there exists infinitely many programs q that purposefully run for much longer than $2^{|q|}$ steps before producing x.

However, thanks to the busy badger renormalisation, this expectation interpretation of the logical depth can be brought to life. I Define the *expected time on the* Ω *clock* as

$$\mathbb{E}(\theta_x) \overset{\text{df}}{=} \sum_{p:\mathcal{U}(p)=x} 2^{-|p|}\theta(p),$$

which can be shown to converge for all x. It suffices to show that it converges for halting programs. Indeed,

$$\mathbb{E}(\theta_{\searrow}) \overset{\text{df}}{=} \sum_{p:\mathcal{U}(p)=\searrow} 2^{-|p|}\theta(p) \leq \sum_{\theta} 2^{-\theta}\theta = 2\,.$$

The inequality comes from reorganizing the sum, and noticing that the mass of programs running in time θ or slower on the Ω clock is less than $2^{-\theta}$, otherwise the value of Ω would be contradicted. This meaningful notion of logical depth as expected running time could perhaps be connected to existing concepts, such as $\mathrm{Depth}_0^{\mathsf{B}}(x)$, which could enhance the justification of its use as the parameter-free depth.

Programs as Ideas

In Ref. [11], Geoffroy Bergeron and I suggested that the notion of emergence could be associated with the existence of strings that display many drops in their structure function, or in their Λ_x profile. Algorithmic models that witness a drop can be thought of a *new idea*, or a new way to explain the data x. Understanding this concept from the time profile $\mathcal{L}_x$ perspective, one finds that those new ideas are equally expressed by programs. The fast but long "`Print` x" program expresses something radically different from the slow but short x^*. In the middle, everything is possible for some strings thanks to "All shapes are possible". In particular, there exists a string that admits a very slow x^* and a very fast program p_{scoop}, of length that exceeds $K(x)$ only by an additive logarithmic term...

Algorithmic Randomness in the Universe.

Preeminent physical theories indicate that the Universe originated in a simple state, and has ever since followed algorithmically simple laws. Through a lengthy computation of 14 billion years on what could be thought of as the most powerful computer of the Universe — the Universe itself — interesting, non-trivial, deep structures emerged. This is the essence of logical depth.

But what superficially appeared as an easier question might in fact remain a puzzle: how can incidental randomness — genuine algorithmic randomness — come about from a simple "computable" Universe? I see two elements of a tentative answer. First, the only kind of such algorithmic randomness that could be generated is halting information.

And it will prosaically arise in time, as any increasing numbers solve ever more halting problems[10].

Second, what we may think to be fragments of disorder, genuine incidental randomness independent of Ω, may in fact only be pieces of antistochasticity. In surface, they seam to be useless noise, but may in fact encode, holographically, the truths about the Universe, *i.e.*, halting information [26]. This holographic encoding of such deep facts may explain what Deutsch [32] refers to as "[o]ne of the most remarkable things about science", namely, "the contrast between the enormous reach and power of our best theories and the precarious, local means by which we create them."

Acknowledgements

My work is supported by Canada's Natural Sciences and Engineering Research Council (NSERC). I am grateful to Charles H. Bennett, Geoffroy Bergeron, Gilles Brassard, Xavier Coiteux-Roy, Samuel Ducharme and Pierre McKenzie for fruitful discussions. I also wish to thank the Institute for Quantum Optics and Quantum Information of Vienna, in particular, Marcus Huber's group, for a warm welcome and inspiring discussions. Last but not least, I am grateful to the veranda, at *Le domaine du pin solitaire*, where this work has been elaborated.

[10] This vision is in sharp contrast with Levin's who does not believe that strings with significant mutual information with the halting problem could exist in the world [56, 75].

4.7. Appendix: Holographic Reconstruction from Time Considerations

I comment briefly on Milovanov's holographic reconstruction understood by time considerations. Consider as before an antistochastic string $z = ab$ of length n and complexity $k \leq n$ and let $|a| = k$. Because of its length, $K(a) \lesssim k$; but also, $K(a) \gtrsim k$. In fact, running a^* and concatenating it to b is one way to compute z, which is of length $K(a) + n - k$ and runs for at most $\mathsf{B}(K(a))$. This contradicts the time profile $\mathcal{L}_z$ unless $K(a) \sim k$. This also means that a^* has busy running time $\sim k$, namely, the same as running time as z^*.
Claim: There are at most $2^{s+O(\log n)}$ programs of length $\leq k$ that halt after $\mathsf{B}(k-s)$ steps[11].

So letting $s = O(\log n)$, z^* (and so z), can be found from an $O(\log n)$ advice if $\mathsf{B}(k)$ — or $\mathcal{H}^{\leq k}$ — is known, because z^* is known to be in the last $2^{O(\log n)}$ halting programs. Therefore, the antistochastic string z becomes simple if the halting problem is solved, which is what a is for. In fact, from a, the logarithmic advice is $K(a)$ permits to find a^*, and by its running time compute $\mathsf{B}(k)$ or $\mathcal{H}^{\leq k}$.

Since antistochastic strings know so much about the halting problem, the halting problem knows so much about them, making them simple! This is the essence of the holographic idea. Any piece of information that solves $\mathcal{H}^{\leq k}$ renders z simple to determine, because one can now start specifying strings from the end of the enumeration. And the particularity of the $\mathcal{L}_z$ profile ensures that any piece of it that is long enough can be use to determine $\mathcal{H}^{\leq k}$ from a logarithmic advice.

[11] This is shown in Ref. [76, Proposition 13]. Otherwise, it can be understood from the closeness between the busy beaver and the busy badger, and that if too many programs are left to halt, the sum M would overshoot Ω.

Chapitre 5

Topics on Quantum Locality

Abstract. It has been 20 years since Deutsch and Hayden demonstrated that quantum systems can be completely described locally — notwithstanding Bell's theorem. More recently, Raymond-Robichaud proposed another approach to the same conclusion. Here, these means of describing quantum systems are shown to be equivalent. Then, they have their cost of description quantified by the dimensionality of their space: The dimension of a single qubit grows exponentially with the size of the total system considered. Finally, the methods are generalized to continuous systems.

> But to admit things not visible to the gross creatures that we are is, in my opinion, to show a decent humility, and not just a lamentable addiction to metaphysics.
>
> — John S. Bell [14] —

5.1. Introduction

It is still a widespread belief that a complete description of a composite entangled quantum system cannot be obtained by descriptions of the parts, if those are expressed independently of what happens to other parts. This apparently holistic feature of entangled quantum states entails violation of Bell inequalities [13, 5] and quantum teleportation [17], which are repeatedly invoked to sanctify the "nonlocal" character of quantum theory. But this widespread belief has been proven false more than twenty years ago by Deutsch and Hayden [35], who by the same token, provided an entirely local explanation of Bell-inequality violations and teleportation.

Descriptions of dynamically isolated — but possibly entangled — systems A and B are *local*[1] if that of A is unaffected by any process system B may undergo, and vice versa. The descriptions are *complete* if they can predict the distributions of any measurement performed on the whole system AB. For instance, if AB is in a pure entangled state $|\Psi\rangle^{AB}$, the reduced density matrices

$$\rho^A = \mathrm{tr}_B|\Psi\rangle\langle\Psi| \qquad \text{and} \qquad \rho^B = \mathrm{tr}_A|\Psi\rangle\langle\Psi|$$

are local but incomplete descriptions. If instead the descriptions of A and B are both taken to be the global wave function $|\Psi\rangle^{AB}$, then one finds a complete but nonlocal account.

Following Gottesman's [49] quantum computation in the Heisenberg picture, Deutsch and Hayden define so-called *descriptors* for individual qubits, which can be intuited to encode the quantum information of a qubit in a Heisenberg-picture-inspired object. Such a mode of description is showed to be both local and complete, hence vindicating the locality of quantum theory. More recently, Raymond-Robichaud has shown that any non-signalling theory with reversible operations can be reformulated in terms of so-called noumenal states, which also satisfy the desirable properties [19]. As a special case of such a non-signalling theory, quantum mechanics also finds noumenal states, as prescribed by Raymond-Robichaud in Ref. [65, Chapter 4].

Mode of description	Local	Complete
Reduced density matrices ρ^A and ρ^B	Yes	No
Global wave function $|\Psi\rangle^{AB}$	No	Yes
DH's descriptors & RR's noumenal states	Yes	Yes

In this paper, equivalences between DH's descriptors, RR's abstract noumenal states and their quantum prescription are established (§5.3). An important drawback of such local descriptions is demonstrated: The dimensionality of the state space of a system as tiny as a qubit scales exponentially with the whole system considered (§5.4). Finally, the formalism is extended to continuous degrees of freedom (§5.5).

[1] After Bell, it has become conventional wisdom to equate locality with a possible explanation by a local hidden variable theory. However, local hidden variables are only one way in which locality can be instantiated [20]. Here, locality is taken in the spirit of Einstein: "the real factual situation of the system S_2 is independent of what is done with the system S_1, which is spatially separated from the former" [69].

5.2. Preliminaries

The DH formalism [35], as well as RR's abstract [19] formalism and its quantum instantiation [65, Chapter 4] are briefly covered in this section. For a more elementary and more detailed introduction to the DH formalism, see the Appendix 5.7.

5.2.1. Deutsch-Hayden's Formalism

Let $\mho$ be a computational network of n qubits. At time 0 the *descriptor* of qubit i is given by

$$\boldsymbol{q}_i(0) = \mathbb{1}^{i-1} \otimes (\sigma_x, \sigma_z) \otimes \mathbb{1}^{n-i}\,,$$

where σ_x and σ_z are the corresponding Pauli matrices. The descriptor is therefore a vector of two[2] components, each of which being an operator on the whole network. Suppose that between the discrete times $s-1$ and s, only one gate is performed, whose matrix representation is denoted G_s. Let $U = G_t \dots G_2 G_1$. The descriptor of qubit i at time t is given by

$$\boldsymbol{q}_i(t) = U^\dagger \boldsymbol{q}_i(0) U\,.$$

The object of n components that encodes the descriptor of each qubit is noted $\boldsymbol{q}(t)$. Alternatively, $\boldsymbol{q}_i(t)$ can be expressed as

$$\boldsymbol{q}_i(t) = \mathsf{U}^\dagger_{G_t}(\boldsymbol{q}(t-1))\boldsymbol{q}_i(t-1)\mathsf{U}_{G_t}(\boldsymbol{q}(t-1))\,,$$

where $\mathsf{U}_{G_t}(\cdot)$ is a fixed operator valued function of some components of $\boldsymbol{q}(t)$ such that $\mathsf{U}_{G_t}(\boldsymbol{q}(0)) = G_t$. In fact, if $U = G_t V$, then

$$\begin{aligned}
\boldsymbol{q}_i(t) &= V^\dagger G_t^\dagger \boldsymbol{q}_i(0) G_t V \\
&= V^\dagger \mathsf{U}^\dagger_{G_t}(\boldsymbol{q}(0)) V V^\dagger \boldsymbol{q}_i(0) V V^\dagger \mathsf{U}_{G_t}(\boldsymbol{q}(0)) V \\
&= \mathsf{U}^\dagger_{G_t}(V^\dagger \boldsymbol{q}(0) V)\boldsymbol{q}_i(t-1)\mathsf{U}_{G_t}(V^\dagger \boldsymbol{q}(0) V) \\
&= \mathsf{U}^\dagger_{G_t}(\boldsymbol{q}(t-1))\boldsymbol{q}_i(t-1)\mathsf{U}_{G_t}(\boldsymbol{q}(t-1)).
\end{aligned}$$

The locality of the descriptors is recognized by the following. If the gate G_t acts only on qubits of the subset $I \subset \{1, 2, \dots, n\}$, then its functional representation U_{G_t} shall only

[2] Deutsch and Hayden originally defined the descriptor with a third component, namely, with σ_y. It is however redundant.

depend on components of $\boldsymbol{q}_k(t-1)$, for $k \in I$. For $j \notin I$, the descriptor $\boldsymbol{q}_j(t-1)$ shall then commute with $U_{G_t}(\boldsymbol{q}(t-1))$, so it will remain unchanged between times $t-1$ and t.

Deutsch and Hayden's descriptors are also complete, in that the expectation value of any observable $\mathcal{O}(t)$ that concerns only qubits of I can be determined by the descriptors $\boldsymbol{q}_k(t)$, with $k \in I$. This can be seen more clearly at time 0, where an observable on the qubits of I is a linear (hermitian) operator that acts non-trivially *only* on the qubits of I. Since any such operator can be generated additively and multiplicatively by the components of $\boldsymbol{q}_k(0)$, with $k \in I$,

$$\mathcal{O}(0) = f_{\mathcal{O}}(\{\boldsymbol{q}_k(0)\}_{k\in I}), \qquad \text{so} \qquad \mathcal{O}(t) = U^\dagger \mathcal{O}(0) U = f_{\mathcal{O}}(\{\boldsymbol{q}_k(t)\}_{k\in I}).$$

5.2.2. Abstract Formalism of Parallel Lives

Systems form a boolean algebra. Specifically, the union and the intersection of systems are systems, and there exist a *whole system* S and an *empty system* $\emptyset$ with respect to which systems can be complemented, *i.e.*, $\bar{A}$ satisfies $\bar{A} \cup A = S$ and $\bar{A} \cap A = \emptyset$.

To each system A is associated a *noumenal state* N^A, a "real state of affairs", from which a *phenomenal state* ρ^A can be determined by an injective function, $\varphi(N^A) = \rho^A$. The phenomenal state encompasses all that can be observed, which may be informationally coarser than the noumenal state. In quantum theory, the phenomenal state boils down to the density matrix of the system, justifying the notation.

To system A is also associated a group of transformations $\mathsf{Op}(A)$ whose elements have an action on both noumenal[3] and phenomenal states. The function φ is promoted to a morphism, since it preserves the group action, namely, for any $V \in \mathsf{Op}(A)$,

$$\varphi(V \cdot N^A) = V * \rho^A,$$

where $\cdot$ and $*$ denote the actions on noumenal and phenomenal states, respectively. The morphism φ also preserves the *tracing out* of systems,

$$\varphi(\mathrm{tr}_B N^{AB}) = \mathrm{tr}_B \rho^{AB},$$

where $\mathrm{tr}_B(\cdot)$ returns a state of system A from that of system AB.

[3] The action is faithful on noumenal states, which means that if $V \cdot N^A = \bar{V} \cdot N^A$ for all N^A, then $V = \bar{V}$.

Evolution and tracing out are merely paralleled by noumenal and phenomenal states, but the whole relevance of introducing noumenal states is to impose that these must be described locally. Raymond-Robichaud makes this locality explicit, in that they impose the existence of a *join product*, noted $\odot$, such that any noumenal state of a joint system AB can be obtained by merging the local descriptions of A and of B,

$$N^{AB} = N^A \odot N^B .$$

If $V \in \mathsf{Op}(A)$ and $W \in \mathsf{Op}(B)$, then the direct product $V \times W$, defined by its action on local noumenal states as

$$(V \times W) \cdot N^{AB} = \left(V \cdot N^A\right) \odot \left(W \cdot N^B\right),$$

is required to be a valid operation on AB. Transformations U and U' on the whole system S are *equivalent with respect to A*, noted $U \sim^A U'$, if they are connected by a transformation that acts trivially on A,

$$U \sim^A U' \iff \exists W \in \mathsf{Op}(\bar{A})\colon\ U' = (\mathbb{1}^A \times W)U . \tag{5.1}$$

In the abstract formalism of Raymond-Robichaud, the noumenal state space associated to system A is defined as the set of equivalence classes, and a particular noumenal state is then

$$N^A \overset{\text{df}}{=} [U]^A .$$

This equivalence class $[U]^A$ encodes what has happened to the whole system S since the beginning, up to evolutions that do not causally concern system A. From such a definition of the noumenal states, evolution by $V \in \mathsf{Op}(A)$, tracing out and merging are defined as

$$V \cdot [U]^A \overset{\text{df}}{=} [V \times \mathbb{1}^{\overline{A}} U]^A ,\ \mathrm{tr}_B [U]^{AB} \overset{\text{df}}{=} [U]^A \text{ and } [U]^A \odot [U]^B \overset{\text{df}}{=} [U]^{AB} . \tag{5.2}$$

Finally, the morphism φ depends upon a reference phenomenal state ρ_0 on system S, and is defined as

$$\varphi([U]^A) \overset{\text{df}}{=} \mathrm{tr}_{\overline{A}}(U * \rho_0) . \tag{5.3}$$

5.2.3. Quantum Formalism of Parallel Lives

Let A be a subsystem of the whole system S, and let $\mathcal{H}^A$ be its Hilbert space, with some basis $\{|i\rangle^A\}$. In the quantum formalism, the noumenal state of system A is defined, not as an equivalence class; rather as an *evolution matrix*,

$$N^A = [\![U]\!]^A \text{, whose matrix elements are } [\![U]\!]^A_{ij} = U^\dagger(|j\rangle\langle i|^A \otimes \mathbb{1}^{\overline{A}})U \,.$$

As in the abstract case, U is the operation that occurred on S between time 0 and time t. The dependence of the evolution matrix on U is only up to the $\sim^A$ equivalence relation, which is defined analogously as in Eq. (5.1). Indeed, if $U' = (\mathbb{1}^A \otimes V)U$,

$$\begin{aligned} [\![U']\!]^A_{ij} &= U'^\dagger(|j\rangle\langle i|^A \otimes \mathbb{1}^{\overline{A}})U' \\ &= U^\dagger(\mathbb{1}^A \otimes V^\dagger)(|j\rangle\langle i|^A \otimes \mathbb{1}^{\overline{A}})(\mathbb{1}^A \otimes V)U \\ &= U^\dagger(|j\rangle\langle i|^A \otimes \mathbb{1}^{\overline{A}})U \\ &= [\![U]\!]_{ij} \,, \end{aligned} \tag{5.4}$$

and one finds the same evolution matrix. The invariance of the evolution matrix within the equivalence class $[\cdot]^A$ is necessary but insufficient to *identify* the evolution matrix with the equivalence class, defined as the noumenal state in the abstract formalism. But Theorem 3 justifies the identification by proving that the equivalence class is uniquely determined by the evolution matrix.

In quantum theory, $\mathsf{Op}(A)$ is the group of unitary transformations $\mathsf{U}(\mathcal{H}^A)$. Let A and B be disjoint systems. Then evolution by $V \in \mathsf{U}(\mathcal{H}^A)$, tracing out and merging are defined as

$$\begin{aligned} \left(V[\![U]\!]^A\right)_{ij} &\overset{\text{df}}{=} \sum_{mn} V_{im}[\![U]\!]^A_{mn}V^\dagger_{nj} \\ \left(\mathrm{tr}_B[\![U]\!]^{AB}\right)_{ij} &\overset{\text{df}}{=} \sum_k [\![U]\!]^{AB}_{ik;jk} \\ \left([\![U]\!]^A \odot [\![U]\!]^B\right)_{ik;jl} &\overset{\text{df}}{=} [\![U]\!]^A_{ij}[\![U]\!]^B_{kl} \,. \end{aligned}$$

The above definitions are quite different from those of the abstract formalism, displayed in Eqns (5.2). Remarkably, these relations instead find their analogues as theorems, derived from the above definitions.

Theorem 1 (Raymond-Robichaud). *Let A and B be disjoint systems and let $V \in \mathrm{U}(\mathcal{H}^A)$. Then*

$$V\llbracket U\rrbracket^A = \llbracket (V \otimes \mathbb{1}^{\overline{A}})U\rrbracket^A, \ \mathrm{tr}_B\llbracket U\rrbracket^{AB} = \llbracket U\rrbracket^A \ \textit{and} \ \llbracket U\rrbracket^A \odot \llbracket U\rrbracket^B = \llbracket U\rrbracket^{AB}.$$

The morphism φ is defined from a fixed reference density matrix ρ_0 as

$$\left(\varphi\llbracket U\rrbracket^A\right)_{ij} \overset{\mathrm{df}}{=} \mathrm{tr}\left(\llbracket U\rrbracket^A_{ij}\rho_0\right). \tag{5.5}$$

Notice that this definition differs from its abstract counterpart, Eq. (5.3), which will again be derived as a theorem. Moreover, the following theorem verifies that the morphism φ intertwines evolution and tracing out, so that these relations are in fact paralleled by noumenal and phenomenal states.

Theorem 2 (Raymond-Robichaud). *Let A and B be disjoint systems and let $V \in \mathrm{U}(\mathcal{H}^A)$. Then*

$$\varphi\llbracket U\rrbracket^A = \mathrm{tr}_{\overline{A}}(U * \rho_0), \ V * \varphi\llbracket U\rrbracket^A = \varphi(V \cdot \llbracket U\rrbracket^A) \ \textit{and} \ \mathrm{tr}_B\varphi\llbracket U\rrbracket^A = \varphi\mathrm{tr}_B\llbracket U\rrbracket^A.$$

One must recall that in quantum theory, the action $*$ of operations on phenomenal states is given by $U * \rho = U\rho U^\dagger$.

5.3. Equivalences

The three approches to quantum locality presented in §5.2 are equivalent in many respects. First the descriptors and the evolution matrices are related by a mere change of operator basis. Second, the quantum formalism of parallel lives can be seen as the instantiation of the abstract one, because the evolution matrices are identified to the equivalence class, at least for qubits.

5.3.1. DH's Descriptor ↔ RR's Evolution Matrix

To establish the equivalence between descriptors and evolution matrices, consider an n-qubit computational network $\mathfrak{N}$, and let $\mathfrak{Q}_k$ denote the k-th qubit. The apparent lack of generality to restrict the considered quantum system to a network of qubits is lifted by their ability to simulate any other quantum system with arbitrary accuracy [31].

At time t, the descriptor of $\mathfrak{Q}_k$ is given by

$$\boldsymbol{q}_k(t) = U^\dagger(\mathbb{1}^{k-1} \otimes \sigma_x \otimes \mathbb{1}^{n-k}, \mathbb{1}^{k-1} \otimes \sigma_z \otimes \mathbb{1}^{n-k})U\,,$$

while its evolution matrix is given by

$$[\![U]\!]_{ij}^{\mathfrak{Q}_k} = U^\dagger(|j\rangle\langle i| \otimes \mathbb{1}^{\overline{\mathfrak{Q}_k}})U\,.$$

In both cases, U is the unitary operator according to which the network has so far evolved, and notwithstanding the different notation, the identity operators are applied on the same subspaces. They can be seen to be informationally equivalent, namely, $[\![U]\!]^{\mathfrak{Q}_k}$ can be computed from $\boldsymbol{q}_k(t)$ and vice versa. In fact, they differ only by a change of operator basis; descriptors are expressed in the Pauli basis, and evolution matrices, in the canonical matrix basis. One should keep in mind that while the descriptor is composed of only two operators, $q_{kx}(t)$ and $q_{kz}(t)$, their multiplicative abilities permit the reconstruction of $q_{ky}(t) = iq_{kx}(t)q_{kz}(t)$. Therefore,

$$\begin{array}{rcl} q_{kx}(t) &=& [\![U]\!]_{12}^{\mathfrak{Q}_k} + [\![U]\!]_{21}^{\mathfrak{Q}_k} \\ q_{kz}(t) &=& [\![U]\!]_{11}^{\mathfrak{Q}_k} - [\![U]\!]_{22}^{\mathfrak{Q}_k} \end{array} \qquad \begin{array}{rcl} [\![U]\!]_{11}^{\mathfrak{Q}_k} &=& \dfrac{\mathbb{1}^{\otimes n} + q_{kz}(t)}{2} \\ [\![U]\!]_{12}^{\mathfrak{Q}_k} &=& \dfrac{q_{kx}(t) - iq_{ky}(t)}{2} \\ [\![U]\!]_{21}^{\mathfrak{Q}_k} &=& \dfrac{q_{kx}(t) + iq_{ky}(t)}{2} \\ [\![U]\!]_{22}^{\mathfrak{Q}_k} &=& \dfrac{\mathbb{1}^{\otimes n} - q_{kz}(t)}{2}\,. \end{array}$$

The connection to observations is also equivalent in both formalisms. Without loss of generality, the reference density matrix ρ_0 can be fixed to $|0\rangle\langle 0|$. In fact, purity can be consecrated in the Church of the larger Hilbert space and from there, altering the global evolution U permits to fix the reference state. The reduced density matrix $\rho(t) = \mathrm{tr}_{\overline{\mathfrak{Q}_k}}(U|0\rangle\langle 0|U^\dagger)$ of qubit $\mathfrak{Q}_k$ at time t can be expressed in the Pauli basis as

$$\rho(t) = \frac{1}{2}\left(\mathbb{1} + \sum_{w\in\{x,y,z\}} p_w(t)\sigma_w\right).$$

From the trace relations of Pauli matrices, the components $p_w(t)$ are

$$p_w(t) = \mathrm{tr}(\rho(t)\sigma_w) = \mathrm{tr}\left(U|0\rangle\langle 0|U^\dagger(\mathbb{1}^{k-1} \otimes \sigma_w \otimes \mathbb{1}^{n-k})\right) = \langle 0|q_{kw}(t)|0\rangle.$$

The second equality from the left comes from that $\rho^A \mapsto \rho^A \otimes \mathbb{1}^B$ is, as a super-operator, the adjoint of $\rho^{AB} \mapsto \mathrm{tr}_B(\rho^{AB})$, and the rightmost equality follows from cyclicality of the trace.

In the evolution matrices framework, one can instead expand the reduced density matrix in its canonical representaiton $\rho(t) = \sum_{ij} \rho_{ij}(t)|i\rangle\langle j|$. The matrix elements can be obtained as

$$\rho_{ij}(t) = \mathrm{tr}(\rho(t)|j\rangle\langle i|) = \mathrm{tr}\left(U|0\rangle\langle 0|U^\dagger(|j\rangle\langle i| \otimes \mathbb{1}^{\overline{\mathfrak{Q}_k}})\right) = \mathrm{tr}\left([\![U]\!]^{\mathfrak{Q}_k}_{ij}|0\rangle\langle 0|\right),$$

consistently with the definition of the morphism, Eq. (5.5).

5.3.2. RR: Abstract → Quantum

The following theorem permits to identify equivalence classes with evolution matrices in the case of qubits.

Theorem 3. *Let $\mathfrak{N}$ be an n-qubit computational network, and let $\mathfrak{Q}_k$ denote the k-th qubit. For all possible evolutions U and U' of $\mathfrak{N}$,*

$$[U]^{\mathfrak{Q}_k} = [U']^{\mathfrak{Q}_k} \iff [\![U]\!]^{\mathfrak{Q}_k} = [\![U']\!]^{\mathfrak{Q}_k}.$$

Proof. The " $\Longrightarrow$ " has already been established by Raymond-Robichaud, and is presented in eq. (5.4) of § 5.2.

Thanks to the DH-RR equivalence, $[\![U]\!]^{\mathfrak{Q}_k}$ can be equivalently represented by

$$\boldsymbol{q}_k(t) = U^\dagger \boldsymbol{q}_k(0) U.$$

To prove the "$\Longleftarrow$", assume $[U]^{\mathfrak{Q}_k} \neq [U']^{\mathfrak{Q}_k}$ and therefore, $U' \neq (\mathbb{1}^{\overline{\mathfrak{Q}_k}} \otimes V)U$, for some V acting on qubit k. Hence, $U' = MU$, for some global operator M, whose functional form $U_M(\boldsymbol{q}(0))$ depends explicitly on terms of $\boldsymbol{q}_k(0)$. But then, if M is thought to occur between time t and t',

$$\begin{aligned} \boldsymbol{q}_k(t') &= U^\dagger M^\dagger \boldsymbol{q}_k(0) M U \\ &= U^\dagger M^\dagger U U^\dagger \boldsymbol{q}_k(0) U U^\dagger M U \\ &= U_M^\dagger(\boldsymbol{q}(t)) \boldsymbol{q}_k(t) U_M(\boldsymbol{q}(t)). \end{aligned}$$

But because of its dependence on $\boldsymbol{q}_k(t)$, $U_M(\boldsymbol{q}(t))$ acts nontrivially on $\boldsymbol{q}_k(t)$ which changes it to a $\boldsymbol{q}_k(t') \neq \boldsymbol{q}_k(t)$, *i.e.*, $[\![U]\!]^{\mathfrak{Q}_k} \neq [\![U']\!]^{\mathfrak{Q}_k}$. □

The previous theorem allows, at least for qubits[4], to unify the definitions of the abstract and the quantum formalisms of parallel lives. The abstract notion of a noumenal state, defined as the equivalence class, can now be realized by the evolution matrix in the quantum setting.

5.4. The Cost of Locality

A standard measure of complexity of an object that can continuously vary is the dimensionality of the space to which it belongs, or the number of degrees of freedom. After the density-matrix space of a qubit is recalled, the descriptor spaces of a single qubit and of a whole network are investigated.

5.4.1. Density-Matrix Space of a Qubit

Consider first the well-known example of the density-matrix space of a single qubit $\mathfrak{Q}_k$ within an n-qubit network $\mathfrak{N}$. In RR's terminology, this is the phenomenal space. The geometric object that characterizes such a state space, notwithstanding the size of the total system to which it belongs, is a unit ball in $\mathbb{R}^3$. This comes from the one-to-one correspondence between the density matrices over a qubit and the points on and inside the Bloch sphere, *i.e.*,

$$\rho = \frac{1}{2}(\mathbb{1} + \boldsymbol{p} \cdot \boldsymbol{\sigma}),$$

where the polarisation vector $\boldsymbol{p} = (p_x, p_y, p_z)$ is constrained by $|\boldsymbol{p}| \leq 1$. The space of density matrices of a qubit ranges along with the range of $\boldsymbol{p}$, which is the unit ball in $\mathbb{R}^3$,

$$\mathsf{Density}^{\mathfrak{Q}_k} \simeq D^3 = \{\boldsymbol{p} \in \mathbb{R}^3 : |\boldsymbol{p}| \leq 1\}.$$

In particular, $\mathrm{Dim}(\mathsf{Density}^{\mathfrak{Q}_k}) = 3$.

[4] The proof could be extended to more general systems, but the analysis for qubits was eased by the DH formalism. For a system A of arbitrary dimension, one can generalize the methods of the DH formalism by constructing a generating set of traceless operators acting on A and $\overline{A}$. This can be achieved with a generalization of Pauli matrices.

5.4.2. Descriptor Space of a Qubit in an n-Qubit Network

How big is the descriptor space — or equivalently, the noumenal space — of a qubit then? Unlike the density-matrix space, the dimension of the descriptor space of a qubit scales (exponentially!) with the size of the whole system Ω to which it belongs. The proof of this assertion involves basic notions of Lie groups, which, in the present context, can be simplified to the special case of a regular hypersurface[5] endowed with a group structure. For more on the topic, see, for instance, Ref. [54].

From the equivalences established in §5.3, a descriptor space of a qubit can be identified to the space of equivalence classes. Define $H \subset \mathrm{U}(2^n)$ the set of operations of the form $\mathbb{1}^{\mathcal{Q}_k} \otimes V$, where $V \in \mathrm{U}(2^{n-1})$ acts on $\overline{\mathcal{Q}_k}$. It is a closed subgroup of $\mathrm{U}(2^n)$, so also a Lie group. Therefore,

$$\mathsf{Descriptor}^{\mathcal{Q}_k \subseteq \Omega} \simeq \mathrm{U}(2^n)/H\,.$$

Let $\bar{k} \neq k$. Denote $\mathrm{CNOT}_{\bar{k}\to k}$ the controlled-not gate in which qubit $\bar{k}$ controls qubit k and denote $N^{\bar{k}}$ the negation gate applied to $\mathcal{Q}_{\bar{k}}$. Then

$$\mathrm{CNOT}_{\bar{k}\to k}(\mathbb{1}^{\mathcal{Q}_k} \otimes N^{\bar{k}})\mathrm{CNOT}_{\bar{k}\to k} \notin H\,,$$

because it does not act trivially on $\mathcal{Q}_k$, in particular, it changes $|00\rangle^{k\bar{k}}$ to $|11\rangle^{k\bar{k}}$. Because CNOT is self-inverse, the above means that H is not a normal subgroup of $\mathrm{U}(2^n)$, and so the quotient $\mathrm{U}(2^n)/H$ is not a group. However, the quotient of Lie groups remains a differential manifold, whose dimension is the difference of the dimensions of the Lie groups involved in the quotient. The group $\mathrm{U}(N)$ has (real) dimension N^2, because it is a hypersurface in $\mathbb{C}^{N^2} \simeq \mathbb{R}^{2N^2}$ subject to the N^2 independent (real) constraints $\sum_j u^*_{ji}u_{jk} = \delta_{ik}$. Since, $H \simeq U(2^{n-1})$, one finds

$$\begin{aligned}\mathrm{Dim}(\mathsf{Descriptor}^{\mathcal{Q}_k \subseteq \Omega}) &= \mathrm{Dim}\mathrm{U}(2^n) - \mathrm{Dim}\mathrm{U}(2^{n-1}) \\ &= 2^{2n} - 2^{2n-2} \\ &= \frac{3}{4} \cdot 2^{2n}\,,\end{aligned}$$

[5] A hypersurface of dimension n is an object defined by m independent constraints in $\mathbb{R}^{n+m}$, $\{y \in \mathbb{R}^{n+m} : F^a(y) = 0\,,\ a = 1,\ldots,m\}$. It is *regular* if the $m \times (n+m)$ matrix with elements $\frac{\partial F^a}{\partial y^i}$ has full rank in all points.

and in particular, the dimension of the descriptor space scales exponentially with the size of the whole system Ω.

Compared to describing the 3-dimensional reduced density matrix of a qubit, if one instead faces the task of describing the *descriptor* of the same qubit, then she must feel like she has the Universe to describe. This is in contradiction with the analysis by Hewitt-Horsman and Vedral [**50**, §3], who claim (in bold font omitted here) that "in general a given state defined by a density matrix has a unique representation in terms of Deutsch-Hayden operators". This statement hinges on a flaw in their analysis: In a nutshell, the number of constraints that determine a descriptor from a density matrix is over counted, so the descriptor should be left under-determined by the density matrix.

Notice that for such an n-qubit network Ω as a whole system, the universal wave function $|\Psi\rangle$, *i.e.*, the Shrödinger state of the whole network, has dimensionality $2^{n+1}-2$. Indeed, the amplitudes are fixed by $2 \cdot 2^n$ real parameters, and the normalization and the irrelevance of a global phase cut down two parameters. Therefore, *the descriptor of a single qubit has larger dimensionality than the Shrödinger state of the whole network — or of the Universe!*

Although the previous statement is surprising, one should not be astounded nor desperate by the exponential scaling of descriptors for single qubits, since it was to be expected. Indeed, the most economical local repartition of the necessary $2^{n+1}-2$ parameters of a complete description of n qubits must still leave $\sim 2^n/n$ parameters in each qubits!

5.4.3. The Universal Descriptor

If the descriptor of a single qubit has larger dimensionality than that of the universal wave function, then how big is the space of universal descriptors? It turns out that it is not much bigger than the qubit descriptor space. The previous analysis can be paralleled, with Ω as the considered system, whose complement is the empty system $\emptyset$. Hence, the subgroup H are the operation of the form $\mathbb{1}^{\Omega} \otimes e^{i\phi}$, which can be identified to U(1). Consequently,

$$\mathsf{Descriptor}^{\Omega} \simeq \mathrm{U}(2^n)/\mathrm{U}(1) \qquad \text{and} \qquad \mathsf{Dim}(\mathsf{Descriptor}^{\Omega}) = 2^{2n}-1\,.$$

Therefore, the universal descriptor is, up to a phase, the unitary operator that occurred on the whole system from time 0 to now. In this case, H is a normal subgroup of $\mathrm{U}(2^n)$, so $\mathsf{Descriptor}^{\Omega}$ keeps a group structure, namely, that of $\mathrm{SU}(2^n)$.

A more pedestrian approach can also be used to establish that a complete description of the whole system entails the knowledge of the evolution U, up to a phase. Indeed, from the descriptors or evolution matrices of each qubit of the network, one can multiplicatively and additively reconstruct $U^\dagger|j\rangle\langle i|U$ for all i and j, where $\{|i\rangle\}_{i=0}^{2^n-1}$ is a basis of $\mathcal{H}^{\Omega}$. The matrix element ℓ, k of $U^\dagger|j\rangle\langle i|U$ is given by

$$\langle\ell|U^\dagger|j\rangle\langle i|U|k\rangle = u_{j\ell}^* u_{ik}\,.$$

By setting $i = j = k = \ell = 0$, one finds $|u_{00}|^2$ and by setting $j = \ell = 0$, but leaving i and k free, one finds $u_{00}^* u_{ik}$ for all i and k. Therefore, up to a phase, U can be computed from $U^\dagger|j\rangle\langle i|U$ for all i and j, which can be computed from $\boldsymbol{q}_i(t)$ or $[\![U]\!]^{\mathfrak{Q}_i}$ for all i.

If the initial state is denoted $|0\rangle$, the universal wave function is obtained (up to a phase), by

$$|\Psi\rangle = U|0\rangle.$$

This corresponds to only one column of the universal descriptor, which is (up to a phase) U, so $U|\phi_0\rangle$ for all possible initial state $|\phi_0\rangle$. If the multiplicity of classical-like terms in Everett's *universal wave function* has prompt some[6] [36] to coin a *Many Worlds Interpretation*, then the multiplicity of Everett's states in a universal descriptor could be thought as many many worlds; namely, as many "many worlds" as there are dimensions in the whole Hilbert space.

5.4.4. What More than the Universal Wave Function?

The many-to-one correspondence between the universal descriptor and the global Schrödinger state (or global density operator) has already been pointed out by Wallace and Timpson [80]. They argued that since the descriptors corresponding to the same Schrödinger state lead to the same observations, they should be equated by some "quantum gauge equivalence". In such a case the description left out boils down again to the usual Schrödinger state, retrieving non-locality. In response, Deutsch [33] attacks the

[6] In his work [38, 39], Everett never reffered to "Many Worlds".

premise and argues that the *dynamics* that has lead to such an actual Schrödinger state, too, may manifest in observations. Indeed, in §5 of his paper, he proposes a way in which one can tell apart different descriptors that yield the same Schrödinger state. Consistently with our identification of the universal descriptor to the evolution operator, his proposal inevitably sums up to network tomography. Raymond-Robichaud, also aware of the *injectivity* of the morphism φ between noumenal and phenomenal states, hold an intermediate standpoint that crops up in their nomenclature. The whole point of their work is to oppose to the Wallace-Timpson identification and authorize — in the name of locality — the existence of noumenal states as elements of reality. They however recognize that different noumenal states may lead to the same observations, encompassed by the phenomenal state.

But what is the extra information that the universal descriptor $\boldsymbol{q}(t)$ gives, that is unobtainable from the universal wave function $|\Psi\rangle$ alone? It can be thought to encode the universal wave function for any possible initial state. In fact, $|\Psi\rangle = U|0\rangle$ is of no use to determine $|\Psi'\rangle = U|0'\rangle$ for a different initial state, with $\langle 0|0'\rangle = 0$. However, $\boldsymbol{q}(t)$ can be used to compute this alternative universal Shrödinger state $|\Psi'\rangle$, or, more in hand with the Heisenberg picture, the expectation $\langle \Psi'|\mathcal{O}|\Psi'\rangle = \langle 0'|U^\dagger \mathcal{O} U|0'\rangle$ of any observable. A computation as such can be done by first defining a unitary operator V such that $V|0\rangle = |0'\rangle$. Recalling that $\mathcal{O}$ can be reconstructed from $\boldsymbol{q}(0)$,

$$\begin{aligned}
\langle 0'|U^\dagger \boldsymbol{q}(0) U|0'\rangle &= \langle 0|V^\dagger U^\dagger \boldsymbol{q}(0) U V|0\rangle \\
&= \langle 0|\mathsf{U}_U^\dagger(\boldsymbol{q}(0'))\,\mathsf{U}_V^\dagger(\boldsymbol{q}(0))\,\boldsymbol{q}(0)\,\mathsf{U}_V(\boldsymbol{q}(0))\,\mathsf{U}_U(\boldsymbol{q}(0'))|0\rangle \\
&= \langle 0|\mathsf{U}_U^\dagger(\boldsymbol{q}(0'))\,\boldsymbol{q}(0')\,\mathsf{U}_U(\boldsymbol{q}(0'))|0\rangle,
\end{aligned}$$

where $0'$ can be thought as an intermediary time delimiting, together with time 0, the application of V. Therefore, since $\boldsymbol{q}(t) = U^\dagger \boldsymbol{q}(0) U$ can be determined by a fixed function of $\boldsymbol{q}(0)$, then $V^\dagger U^\dagger \boldsymbol{q}(0) U V$ is determined by the same function, but instead evaluated on argument $\boldsymbol{q}(0')$.

This puts in evidence a particular feature of the DH formalism, namely, it enables the evolution of the descriptors in both directions in time, simultaneously. On the one hand, adding a gate at the end of the network affects the outer shell, that is to say, the function that determines $\boldsymbol{q}(t+1)$ from $\boldsymbol{q}(0)$ will differ from that of $\boldsymbol{q}(t)$. On the other hand, supplementing a gate at the beginning of the network changes the inner shell: The defining function of $\boldsymbol{q}(t)$ remains the same, but it is instead applied to the argument $\boldsymbol{q}(0')$.

5.5. Continuous Systems

Evolution matrices can naturally be extended to locally describe quantum systems of continuous degrees of freedom. The mathematical structures required to formalize the approach are those of Dirac calculus, once made mathematically meaningful by Schwartz' distribution theory [70]. For a concise presentation, see Ref. [7, p.28].

Consider a system A with a continuous one dimensional observable (*e.g.*, the position of a particle). Associated to this system is a rigged Hilbert space admitting a Dirac-orthonormal basis $\{|x\rangle\}_{x\in\mathbb{R}}$, where

$$\langle x|x'\rangle = \delta(x-x') \qquad \text{and} \qquad \int_{\mathbb{R}} |x\rangle\langle x| = \mathbb{1}\,.$$

The wave function can then be represented spatially by $\psi(x) = \langle x|\psi\rangle$. The evolution matrix associated to the system A is a "continuous matrix[7]" whose matrix elements are given by

$$[\![U]\!]^A_{xy} = U^\dagger\Big(|y\rangle\langle x|\otimes \mathbb{1}^{\overline{A}}\Big)U\,.$$

Here again, U is the evolution that the whole system has undergone, which could have been represented by any other $(\mathbb{1}^A\otimes V)U$. Let A and B be disjoint systems of a continuous one-dimensional observable. Analogously as in §5.2.3, evolution by $V\in \mathrm{U}(\mathcal{H}^A)$, tracing out and merging are defined as

$$\begin{aligned}
\Big(V[\![U]\!]^A\Big)_{xy} &\overset{\text{df}}{=} \int_{\mathbb{R}^2} dx'dy' V_{xx'}[\![U]\!]^A_{x'y'}V^\dagger_{y'y}\\
\Big(\mathrm{tr}_B[\![U]\!]^{AB}\Big)_{xy} &\overset{\text{df}}{=} \int_{\mathbb{R}} dz[\![U]\!]^{AB}_{xz;yz}\\
\Big([\![U]\!]^A\odot[\![U]\!]^B\Big)_{x_Ax_B;y_Ay_B} &\overset{\text{df}}{=} [\![U]\!]^A_{x_Ay_A}[\![U]\!]^B_{x_By_B}\,.
\end{aligned}$$

With those definitions in hand, the analogue of Theorem 1 holds.

Theorem 4. *Let A and B be disjoint systems of a continuous observable and let $V\in U(\mathcal{H}^A)$.*

$$V[\![U]\!]^A \;=\; [\![(V\otimes\mathbb{1}^{\overline{A}})U]\!]^A$$

[7] An object M as such is in fact a sesquilinear form on test functions, which maps f and g to $\int_{\mathbb{R}^2} dxdyM_{xy}f^*(x)g(y)$.

$$\begin{aligned} \mathrm{tr}_B[\![U]\!]^{AB} &= [\![U]\!]^A \\ [\![U]\!]^A \odot [\![U]\!]^B &= [\![U]\!]^{AB}. \end{aligned}$$

For a fixed reference density matrix ρ_0, the morphism φ is defined as

$$\left(\varphi[\![U]\!]^A\right)_{xy} \overset{\mathrm{df}}{=} \mathrm{tr}\left([\![U]\!]^A_{xy}\rho_0\right),$$

and Theorem 2 also generalizes to continuous systems.

Theorem 5. *Let A and B be disjoint systems of a continuous variable and and let $V \in \mathrm{U}(\mathcal{H}^A)$. Then*

$$\varphi[\![U]\!]^A = \mathrm{tr}_{\overline{A}}(U * \rho_0),\ V * \varphi[\![U]\!]^A = \varphi(V[\![U]\!]^A) \textit{ and } \mathrm{tr}_B\varphi[\![U]\!]^A = \varphi\mathrm{tr}_B[\![U]\!]^A.$$

The proofs of theorems 4 and 5 are relegated to Appendix 5.8.

5.6. Conclusions

Deutsch and Hayden conclude their paper with a beautiful analogy that compares their descriptor obtained in the Heisenberg picture to the usual representation of a quantum state framed in the Schrödinger picture:

> *The relationship between the two pictures is somewhat analogous to that between any descriptive piece of information, such as a text or a digitized image, and an algorithmically compressed version of the same information that eliminates redundancy to achieve a more compact representation. If the compression algorithm used is not 'lossy', then, considered as a description of the original data, the two versions are mathematically equivalent. However, the elimination of redundancy results in strong interdependence between the elements of the compressed description so that, for instance, a localized change in the original data can result in changes all over the compressed version, so that a particular character or pixel from the original is not necessarily located at any particular position in the compressed version. Nevertheless, it would be a serious error to conclude that this 'holistic' property of the compressed description expresses any analogous property in the original text or image, or of course in the reality that they refer to.*

The underdetermination of the descriptor by the Schrödinger state renders the "compression algorithm" lossy. But the analogy does not collapse; the usual representation of a quantum state may now exhibit holistic features because of its compactness *or* because of its lost information.

As discussed in §5.4.4, the lost information is about the various other dynamics of the network, would it have been initialized differently. In quantum information theory, qubits initialized in a state $|0\rangle$ are taken as a free entity; but how does one really get such an initialized qubit in a unitary quantum realm? This may be referred to as the *preparation problem*, dual to the measurement problem. A parsimonious solution should provide a mechanism that explains, from within unitary quantum theory, why computations can be done *as if* the state really was $|0\rangle$. Such an explanation would rely on decoherence arguments, and in the larger unitary scheme, not only $|0\rangle$ should go through the whole network, perhaps justifying the need for more dynamics.

The complexity of the descriptor was investigated in §5.4 through the dimensionality of its space, well motivated in physics. However, a computer theoretic approach may regard as the complexity cost of a descriptor its difficulty in time, in space or in program size to produce it. An investigation as such should be hand in hand with circuit complexity, since the whole descriptor is but a compact representation of the operator representing its generating circuit. Perhaps, also, a new insight into quantum Kolmogorov complexity could be provided in the DH formalism.

If one is willing to pay Everett's price, and accepts that the n-qubit universe is encoded in a point $|\Psi\rangle$ moving in $2^{n+1} - 2$ dimensions, then one should without regrets square this number up to $2^{2n} - 1$ and instead use the universal descriptor for an entirely local story. One then faces the surprising consequence that more than 3/4 of the whole dimensionality resides in each qubit. Most of this information is locally inaccessible; it accounts for common histories among qubits, keeping track of whom is entangled with whom. The consequence becomes more digestible when one appreciates how entangled the universe really is. And before backing off from the implications of a well-motivated paradigm shift, one reminds Wallace's advice [79] : "The moral is clear: our intuitions as to what is 'unreasonable' or 'absurd' were formed to aid our ancestors scratching a living on the savannahs of Africa, and the Universe is not obliged to conform to them".

Looping the loop with whom we started, Bell also stated [14] that "Either the wave function, as given by the Shrödinger equation, is not everything, or it is not right". It is so far right, but not everything; and completing it by the universal descriptor is perhaps what Einstein Podolsky and Rosen [37] were looking for.

5.7. Appendix: Introduction to the DH Formalism

When bold foundational statements such as those established by Deutsch and Hayden (*cf.* Section 5.1) collect a mere 165 citations in 20 years, it is perhaps because a large portion of the community of quantum foundations is unaware of their contribution, or does not understand it properly. This appendix is an elementary introduction to the DH formalism. It covers Sections 2 and 3 of Ref. [35] in much more length, providing examples of calculations and explanations from different standpoints. It is aimed both for experts and non-experts in quantum theory: A reader with introductory knowledge in quantum information theory, with or without a physics background, should understand this text.

5.7.1. A Question of Picture

In quantum theory, computations leading to measurable quantities all take the same form: They are expected values of some observables. An observable $\mathcal{O}$ is represented by a hermitian operator which admits a spectral decomposition

$$\mathcal{O} = \sum_i \lambda_i \Pi_i \,,$$

where $\lambda_i \in \mathbb{R}$ are the eigenvalues corresponding to the measurement outcomes and the Π_i are the corresponding projectors on the eigensubspaces. If the system is in state $|\psi\rangle$, the expected value of such an observable is given by $\langle\psi|\mathcal{O}|\psi\rangle$, since

$$\langle\psi|\mathcal{O}|\psi\rangle = \langle\psi|\sum_i \lambda_i \Pi_i|\psi\rangle = \sum_i \langle\psi|\Pi_i|\psi\rangle\lambda_i = \sum_i p_i \lambda_i \,,$$

where p_i is the probability of measuring outcome λ_i. This type of computation is routine for physicists, but quantum information scientists usually compute probabilities of measurement outcomes. An n-qubit network in the state

$$\sum_{j=0}^{2^n-1} \alpha_j |j\rangle$$

has a probability $|\alpha_l|^2$ to return the classical value "l". But

$$|\alpha_l|^2 = \langle\psi||l\rangle\langle l||\psi\rangle$$

is nothing but the expectation value of the observable $|l\rangle\langle l|$.

In general $|\psi\rangle$ could be a complex state that comes from a large network applied to the initial state $|0\rangle$, in some fixed basis. Hence, if U is the unitary operator representing the network,

$$|\psi\rangle = U|0\rangle.$$

Therefore, the computations carried to predict statistical properties of the quantities measured in the laboratory all have the form

$$\langle 0|U^\dagger \mathcal{O} U|0\rangle, \tag{5.6}$$

where $|0\rangle$ is the initial state, U is the unitary evolution and $\mathcal{O}$ is the observable.

The *Shrödinger picture* is about viewing the sandwich Equation (5.6) as if the bread evolves and the meat stays constant, namely,

$$\left(\langle 0|U^\dagger\right) \mathcal{O} \left(U|0\rangle\right).$$

With such a viewpoint, the initial state $|0\rangle$ evolves to the final state $|\psi\rangle = U|0\rangle$ and the observable $\mathcal{O}$ remains constant.

The *Heisenberg picture* is about regarding the sandwich equation as if the meat evolves but the bread remains constant,

$$\langle 0|\left(U^\dagger \mathcal{O} U\right)|0\rangle. \tag{5.7}$$

In this picture, the state vector remains fixed to $|0\rangle$ but the observable $\mathcal{O}$ evolves to $U^\dagger \mathcal{O} U$. Therefore, in the Heisenberg picture, the term ‘state’, which refers to a quantity that is fixed to $|0\rangle$ becomes a misnomer. For this reason, it will be referred to as the reference vector. Deutsch and Hayden’s descriptors come from encoding the information of the quantum system into evolving observables, as if one tries to define a “Heisenberg state”.

5.7.2. Tracking Observables

In the Heisenberg picture, a quantum system shall no longer be described by its Schrödinger state, but rather by an object that encodes the information about *all* the evolved observables on the system. Luckily, observables are linear operators and so form a vector space. Since the evolution $\mathcal{O} \to U^\dagger \mathcal{O} U$ is linear, one does not need to track the evolution of infinitely many observables: *Only a basis* of the linear operators suffices. Indeed, if $\mathcal{O} = \sum_j a_j B_j$, then $U^\dagger \mathcal{O} U = \sum_j a_j U^\dagger B_j U$.

5.7.2.1. *The Descriptor of a* 1*-Qubit Network*

In the case of a singe qubit, the Pauli matrices together with the identity,

$$\boldsymbol{\sigma} = (\sigma_x, \sigma_y, \sigma_z) = \left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right) \text{ and } \sigma_0 = \mathbb{1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

form a basis of any 2×2 matrices, if the linear combinaisons are taken over complex numbers. Following the evolution of $\mathbb{1}$ is trivial, $U^\dagger \mathbb{1} U = \mathbb{1}$, so it can be neglected and one only follows the evolution of $\boldsymbol{\sigma}$. Hence, in the Heisenberg picture, a qubit is represented by a *descriptor* $\boldsymbol{q}(t) = U^\dagger \boldsymbol{\sigma} U$, where U is the unitary operator that represents the evolution undergone by quantum network between time 0 and time t.

Example 6. *Describe $|+\rangle$ in the Heisenberg picture.*

Solution : One takes the initial reference vector to be fixed to $|0\rangle$. In the Shrödinger picture, $|+\rangle = H|0\rangle$, where H is the Hadamard gate so the descriptor is given by

$$H^\dagger \boldsymbol{\sigma} H = H(\sigma_x, \sigma_y, \sigma_z) H = (\sigma_z, -\sigma_y, \sigma_x).$$

The descriptor is not uniquely determined by the Shrödinger state $|+\rangle$, since any other unitary transformation U such that $|+\rangle = U|0\rangle$ can be taken instead of H. This underdetermination is explored in more details in Section 5.4.

5.7.2.2. *Descriptors of an n-Qubit Network*

A natural basis to the space of all operators on n qubits is the product of Pauli operators, namely

$$\mathcal{B}^n \equiv \{\sigma_{\mu_1} \otimes \sigma_{\mu_2} \otimes \ldots \sigma_{\mu_n} : \; \mu_i \in \{0, x, y, z\}\}.$$

There are 4^n such matrices, which are linearly independent and hence they form a basis of the $2^n \times 2^n = 4^n$ dimensional complex vector space of linear operators on n-qubits.

This means that if one knows how each observable of the basis evolves by the action of some unitary operator U,

$$\sigma_{\mu_1} \otimes \sigma_{\mu_2} \otimes \ldots \sigma_{\mu_n} \to U^\dagger \sigma_{\mu_1} \otimes \sigma_{\mu_2} \otimes \ldots \sigma_{\mu_n} U, \qquad \mu_i \in \{0, x, y, z\},$$

then we know, by linearity, how each observable evolves.

5.7.2.3. *DH's Shortcut*

In the case of n interacting qubits of some quantum computational network $\mho$, Deutsch and Hayden suggest to track the set of observables

$$\boldsymbol{q}_i(0) = \mathbb{1}^{i-1} \otimes \boldsymbol{\sigma} \otimes \mathbb{1}^{n-i}, \qquad i = 1, \dots, n, \tag{5.8}$$

where $\mathbb{1}^k$ stands for the tensor product of k copies of the identity. Note that for each i, $\boldsymbol{q}_i(0)$ has 3 components. The n-tuple whose components are the $\boldsymbol{q}_i(0)$ is noted $\boldsymbol{q}(0)$. Bold quantities are vectors, so one writes $\boldsymbol{q}_i(0)$, but $q_{ix}(0)$. The vector $\boldsymbol{q}(0)$ represents the initial observables, namely, those at time $t = 0$, whence the notation.

Importantly, note that $\boldsymbol{q}(0)$ contains much fewer components than $\mathcal{B}^n$ contains elements. In fact, instead of tracking the 4^n operators of $\mathcal{B}^n$ only $3n$ are suggested here. The reason is that these $3n$ operators have a multiplicative structure that allows to generate any of the 4^n basis operators. Moreover, this multiplicative structure is preserved by the evolution U, namely, if q and $\bar{q}$ are any operator,

$$q\bar{q} \to (q\bar{q})' = U^\dagger q\bar{q}U = U^\dagger qUU^\dagger\bar{q}U = q'\bar{q}'.$$

Remark 7. *The operators of* $\boldsymbol{q}(0)$ *satisfy the* $\mathfrak{su}(2)^{\otimes n}$ *algebra, namely*

$$\begin{array}{rcll} [q_{iw}(0), q_{jw'}(0)] & = & 0 & (i \neq j \textit{ and } \forall w, w') \\ q_{ix}(0)q_{iy}(0) & = & iq_{iz}(0) & (\textit{and cyclic permutations}) \\ q_{iw}(0)^2 & = & \mathbb{1} & (\forall w). \end{array} \tag{5.9}$$

5.7.2.4. *One more Shortcut*

Following Gottesman [49], the generating tuple $\boldsymbol{q}(0)$ could be reduced to $2n$ elements by noticing a redundancy due to the $\mathfrak{su}(2)^{\otimes n}$ algebra. In fact, only two of the three $(q_{ix}(0), q_{iy}(0), q_{iz}(0))$ operators are required, for any i, since the case operator is obtained by the product of the selected two. In what follows, the notation will not be modified, but one will happily use this shortcut to avoid tracking the observables $q_{iy}(0)$, since $q_{iy}(0) = -iq_{ix}(0)q_{iz}(0)$.

Summing this up, knowing the evolution of the $2n$ observables of $\boldsymbol{q}(0)$ (without the $q_{iy}(0)$) allows to infer, by group multiplication, the evolution of the 4^n observables of $\mathcal{B}_n$, which allows to infer, by linearity, the evolution of any observable.

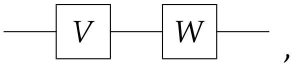
In this case, the descriptor of qubit i at time t is given by

$$\boldsymbol{q}_i(t) = U^\dagger \boldsymbol{q}_i(0) U\,, \tag{EVO 1}$$

where U is the unitary operator that represents the evolution undergone by quantum network between time 0 and time t.

5.7.3. Evolution from the future?!

Although $\mathcal{O} \to U^\dagger \mathcal{O} U$ looks like a completely fine way in which observables should evolve, when U is broken down into different gates, for instance $U = WV$, one finds that the observables evolve in the wrong order! In fact, WV means that V is done before W, or diagrammatically,



but the observable evolves as

$$\mathcal{O} \to V^\dagger W^\dagger \mathcal{O} W V\,, \tag{5.10}$$

i.e., W is applied first, then V. In a computational network, the evolution of observables then occurs from the last gate of the network to the first, which is completely unnatural and in most cases inconvenient, since the network needs to be final before computing anything.

The way out of this conundrum is to notice that inasmuch as observables $\mathcal{O}$ are linear operators generated by some set $\boldsymbol{q}(0)$ of operators, the evolution operators U are too. They are generated multiplicatively and linearly by the same set $\boldsymbol{q}(0)$, since questions of hermicity versus unitarity did not arise.

For a fixed gate with matrix representation G, its generation by $\boldsymbol{q}(0)$ defines a function $U_G(\cdot)$ through

$$G = U_G(\boldsymbol{q}(0))\,. \tag{5.11}$$

The function $U_G(\cdot)$ takes value in unitary operators and will be referred to as the *functional representation* of the gate G. Its functionality encodes the multiplicative and linear generation of G by the elements of $\boldsymbol{q}(0)$. For instance, the familiar negation and Hadamard gates are described by

$$N = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \sigma_x = q_x(0) \qquad \text{and} \qquad H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{q_x(0) + q_z(0)}{\sqrt{2}},$$

so their functional representations are

$$U_N(\boldsymbol{q}(0)) = q_x(0) \qquad \text{and} \qquad U_H(\boldsymbol{q}(0)) = \frac{q_x(0) + q_z(0)}{\sqrt{2}}.$$

The clockwise rotation of a state vector in the $|0\rangle$ & $|1\rangle$ plane[8] is described by

$$R_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} = \cos\theta \mathbb{1} + i \sin\theta \sigma_y = \cos\theta q_x(0)^2 - \sin\theta q_x(0) q_z(0),$$

which defines its functional representation $U_{R_\theta}(\cdot)$.

Now, when $\boldsymbol{q}(t)$ varies with t, the matrix representation $U_G(\boldsymbol{q}(t))$ also varies, but it is the fixed functionality that plays a role in Heisenberg computations.

5.7.3.1. *Back in order!*

Since the usual matrix representation of a gate V is expressed by $U_V(\boldsymbol{q}(0))$, then if V is the first gate of the quantum network, by Equation (EVO 1),

$$\boldsymbol{q}_i(1) = U_V^\dagger(\boldsymbol{q}(0)) \boldsymbol{q}_i(0) U_V(\boldsymbol{q}(0)). \tag{5.12}$$

The apparently reversed ordered evolution of Equation (5.10) can then be transformed back in the right order:

$$\begin{aligned} V^\dagger W^\dagger \mathcal{O} W V &= U_V^\dagger(\boldsymbol{q}(0)) U_W^\dagger(\boldsymbol{q}(0)) \mathcal{O} U_W(\boldsymbol{q}(0)) U_V(\boldsymbol{q}(0)) \\ &= U_V^\dagger(\boldsymbol{q}(0)) U_W^\dagger(\boldsymbol{q}(0)) U_V(\boldsymbol{q}(0)) U_V^\dagger(\boldsymbol{q}(0)) \mathcal{O} U_V(\boldsymbol{q}(0)) U_V^\dagger(\boldsymbol{q}(0)) \end{aligned}$$

[8] Note that this operation represents the rotation of a polarized photon, but not exactly that of the spin of an electron. The reason for this is that a $\pi/2$ rotation of a photon takes the horizontal polarization $|\leftrightarrow\rangle \equiv |0\rangle$ to the vertical polarization $|\updownarrow\rangle \equiv |1\rangle$. However, the spin of an electron needs a π rotation to take the $|\uparrow_z\rangle \equiv |0\rangle$ to $|\downarrow_z\rangle \equiv |1\rangle$. Such a rotation is better represented on the Bloch sphere and shall be discussed in section 5.7.7.1.

$$U_W(\boldsymbol{q}(0))U_V(\boldsymbol{q}(0))$$
$$= \quad U_W^\dagger(\boldsymbol{q}(1))U_V^\dagger(\boldsymbol{q}(0))\mathcal{O}U_V(\boldsymbol{q}(0))U_W(\boldsymbol{q}(1)).$$

Where the last equation, namely that $U_W(\boldsymbol{q}(1)) = U_V^\dagger(\boldsymbol{q}(0))U_W(\boldsymbol{q}(0))U_V(\boldsymbol{q}(0))$, comes from the following. Since $U_W(\boldsymbol{q}(0))$ is some fonction of the components of $\boldsymbol{q}(0)$, when it is sandwiched between $U_V^\dagger(\boldsymbol{q}(0))$ and $U_V(\boldsymbol{q}(0))$, every term containing some $q_{iw}(0)$ gets transformed to its corresponding $q_{iw}(1)$. Terms of $U_W(\boldsymbol{q}(0))$ that contain products $q_{iw}(0)q_{jw'}(0)$ need extra bread in the middle of the sandwich, *i.e.*

$$U_V^\dagger(\boldsymbol{q}(0))q_{iw}(0)U_V(\boldsymbol{q}(0))U_V^\dagger(\boldsymbol{q}(0))q_{jw'}(0)U_V(\boldsymbol{q}(0)), \tag{5.13}$$

yielding $q_{iw}(1)q_{jw'}(1)$.

Iterating the argument, a quantum network of many gates G_1, G_2, $\ldots$, G_N has its observables tracked in two possible ways.

- With the usual fixed matrix representation of unitary operators that act in the wrong order on observables (but in the right order if they were to act in the Shrödinger picture),
$$G_1^\dagger G_2^\dagger \ldots G_N^\dagger \mathcal{O} G_N \ldots G_2 G_1$$
- With the operators defined as a fixed function of the generating set $\boldsymbol{q}(t)$ which act in the right order,

$$U_{G_N}^\dagger(\boldsymbol{q}(N-1))\ldots U_{G_2}^\dagger(\boldsymbol{q}(1))U_{G_1}^\dagger(\boldsymbol{q}(0))\,\mathcal{O}\,U_{G_1}(\boldsymbol{q}(0))U_{G_2}(\boldsymbol{q}(1))\ldots U_{G_N}(\boldsymbol{q}(N-1))$$

The later approach is preferred to perform computations in the Heisenberg picture.

5.7.4. Another Evolution Equation

Deutsch and Hayden do not pass by Equation (EVO 1) to evolve the descriptor from time 0 to t. Instead, the descriptor is claimed to evolve iteratively as

$$\boldsymbol{q}_i(t+1) = U_W^\dagger(\boldsymbol{q}(t))\boldsymbol{q}_i(t)U_W^\dagger(\boldsymbol{q}(t)), \tag{EVO 2}$$

where W is the gate performed on the network between time t and $t+1$. However, such an iterative evolution is equivalent to the one prescribed by Equation (EVO 1).

(EVO 1) $\implies$ (EVO 2). Let V be the unitary operator representing the evolution of the network between time 0 and time t.

$$\begin{aligned}
\boldsymbol{q}_i(t+1) &= (WV)^{\dagger}\boldsymbol{q}_i(0)WV \\
&= V^{\dagger}U_W^{\dagger}(\boldsymbol{q}(0))VV^{\dagger}\boldsymbol{q}_i(0)VV^{\dagger}U_W(\boldsymbol{q}(0))V \\
&= U_W^{\dagger}(\boldsymbol{q}(t))\boldsymbol{q}_i(t)U_W(\boldsymbol{q}(t)).
\end{aligned}$$

(EVO 2) $\implies$ (EVO 1). Let G_s be the gate that occurs between time $s-1$ and s. The base of the induction is easily verified

$$\begin{aligned}
\boldsymbol{q}_i(1) &= U_{G_1}^{\dagger}(\boldsymbol{q}(0))\boldsymbol{q}_i(0)U_{G_1}(\boldsymbol{q}(0)) \\
&= G_1^{\dagger}\boldsymbol{q}_i(0)G_1\,,
\end{aligned}$$

and with induction hypothesis

$$\boldsymbol{q}_i(t-1) = G_1^{\dagger}G_2^{\dagger}\dots G_{t-1}^{\dagger}\boldsymbol{q}(0)G_{t-1}\dots G_2G_1\,,$$

one finds

$$\begin{aligned}
\boldsymbol{q}_i(t) &= U_{G_t}^{\dagger}(\boldsymbol{q}(t-1))\boldsymbol{q}_i(t-1)U_{G_t}^{\dagger}(\boldsymbol{q}(t-1)) \\
&= \dots\boldsymbol{q}_i(0)G_{t-1}\dots G_2G_1U_{G_t}^{\dagger}(\boldsymbol{q}(t-1)) \\
&= \dots\boldsymbol{q}_i(0)G_{t-1}\dots G_2G_1U_{G_t}^{\dagger}\left(G_1^{\dagger}G_2^{\dagger}\dots G_{t-1}^{\dagger}\boldsymbol{q}(0)G_{t-1}\dots G_2G_1\right) \\
&= \dots\boldsymbol{q}_i(0)G_{t-1}\dots G_2G_1G_1^{\dagger}G_2^{\dagger}\dots G_{t-1}^{\dagger}U_{G_t}^{\dagger}(\boldsymbol{q}(0))\,G_{t-1}\dots G_2G_1 \\
&= \dots\boldsymbol{q}_i(0)G_tG_{t-1}\dots G_2G_1\,.
\end{aligned}$$

The fourth line is obtained from the third by a similar argument as in Eq. (5.13). For conciseness, the left of $\boldsymbol{q}_i(0)$ has been omitted since it has a symmetric behaviour as what happens to the right of it.

5.7.5. Not its matrix rep, but its action!

In the Shrödinger picture, the state $|\psi(t)\rangle$ at time t can be computed by the action of the gates of the network on $|\psi(0)\rangle$. The computation of the descriptor $\boldsymbol{q}(t)$ at time t can also conveniently be computed form the action of the gates. However, it is not achieved by matrix multiplication, rather, through the functional representation of the gates and the relations $\mathfrak{su}(2)^{\otimes n}$ algebra.

Remark 8. *Even if $\boldsymbol{q}(t)$ loses its initial tensor product form of Equation (5.8), it still satisfies the $\mathfrak{su}(2)^{\otimes n}$ algebra (Relations (5.9)):*

$$\begin{aligned}
[q_{iw}(t), q_{jw'}(t)] &= q_{iw}(t)q_{jw'}(t) - q_{jw'}(t)q_{iw}(t) \\
&= U^\dagger q_{iw}(0)UU^\dagger q_{jw'}(0)U - U^\dagger q_{jw'}(0)UU^\dagger q_{iw}(0)U \\
&= U^\dagger q_{iw}(0)q_{jw'}(0)U - U^\dagger q_{jw'}(0)q_{iw}(0)U \\
&= U^\dagger [q_{iw}(0), q_{jw'}(0)]U \\
&= 0 \qquad\qquad (i \neq j \textit{ and } \forall w, w')
\end{aligned}$$

$$\begin{aligned}
q_{ix}(t)q_{iy}(t) &= U^\dagger q_{ix}(0)UU^\dagger q_{iy}(0)U \\
&= U^\dagger q_{ix}(0)q_{iy}(0)U \\
&= U^\dagger i q_{iz}(0)U \\
&= i q_{iz}(t) \qquad\qquad (\textit{and cyclic permutations})
\end{aligned}$$

$$\begin{aligned}
q_{iw}(t)^2 &= U^\dagger q_{iw}(0)UU^\dagger q_{iw}(0)U \\
&= U^\dagger q_{iw}(0)q_{iw}(0)U \\
&= U^\dagger \mathbb{1} U \\
&= \mathbb{1} \qquad\qquad (\forall w).
\end{aligned}$$

Let W be the gate performed between time t and time $t+1$. For each i, its action on $\boldsymbol{q}_i(t)$ is

$$\begin{aligned}
W : \boldsymbol{q}_i(t) \rightarrow \boldsymbol{q}_i(t+1) &= U_W^\dagger(\boldsymbol{q}(t))\boldsymbol{q}_i(t)U_W(\boldsymbol{q}(t)) \\
&= U_W^\dagger(\boldsymbol{q}(t))(q_{ix}(t), q_{iz}(t))U_W(\boldsymbol{q}(t)).
\end{aligned}$$

For a generic gate, the updating of $\boldsymbol{q}_i(t)$ to $\boldsymbol{q}_i(t+1)$ requires $2n$ sandwich-like calculations. However, if W acts on only two[9] qubits (*e.g.*, qubits j and k), it reduces to only 4

[9] Universal gate sets can be formed from gates acting on no more than two qubits, for instance, the CNOT supplemented by arbitrary unary gates.

such calculations. Indeed, the linear transformation W (one can think of its matrix representation) acts as the identity on all product spaces that concerns not qubits j and k. Therefore, the functional representation of the gate, defined by $U_W(\boldsymbol{q}(0)) = W$, can only depend on $q_{jx}(t)$, $q_{jz}(t)$, $q_{kx}(t)$ and $q_{kz}(t)$. Because of the preserved algebraic relations, particularly $[q_l(t), q_j(t)] = 0 = [q_l(t), q_k(t)]$, the update of the descriptor $\boldsymbol{q}_l(t)$ is trivial for any qubit different than qubit j or k. The computation is therefore majorly enlightened, and noted

$$\begin{aligned} W : \left\{ \begin{array}{c} \boldsymbol{q}_j(t+1) \\ \boldsymbol{q}_k(t+1) \end{array} \right\} &= U_W^{\dagger}(\boldsymbol{q}(t)) \left\{ \begin{array}{c} \boldsymbol{q}_j(t) \\ \boldsymbol{q}_k(t) \end{array} \right\} U_W(\boldsymbol{q}(t)) \\ &= \left\{ \begin{array}{c} U_W^{\dagger}(\boldsymbol{q}(t))(q_{jx}(t), q_{jz}(t)) U_W(\boldsymbol{q}(t)) \\ U_W^{\dagger}(\boldsymbol{q}(t))(q_{kx}(t), q_{kz}(t)) U_W(\boldsymbol{q}(t)) \end{array} \right\}. \end{aligned}$$

5.7.6. Examples

Let H_i denote the Hadamard gate H performed on the i-th qubit.

$$U_{H_i}(\boldsymbol{q}(t)) = \frac{q_{ix}(t) + q_{iz}(t)}{\sqrt{2}}.$$

The action on descriptor $\boldsymbol{q}_i$ is then

$$\begin{aligned} H_i : (q_{ix}(t), q_{iz}(t)) &\rightarrow (q_{ix}(t+1), q_{iz}(t+1)) \\ &= \frac{q_{ix}(t) + q_{iz}(t)}{\sqrt{2}} (q_{ix}(t), q_{iz}(t)) \frac{q_{ix}(t) + q_{iz}(t)}{\sqrt{2}} \\ &= \frac{1}{2}(q_{ix} + q_{iz} + q_{iz} - q_{ix}, -q_{iz} + q_{ix} + q_{ix} + q_{iz}) \\ &= (q_{iz}, q_{ix}). \end{aligned}$$

When the context does not require it, "(t)" can be omitted and one notes $H_i : (q_{ix}, q_{iz}) \rightarrow (q_{iz}, q_{ix})$. And when not specified, all the other $\boldsymbol{q}_k$ with $k \neq i$ remain unchanged by the action by H_i.

The negation gate N on qubit i has $U_{N_i}(\boldsymbol{q}(t)) = q_{ix}(t)$ and so

$$N_i : (q_{ix}, q_{iz}) \rightarrow (q_{ix}, -q_{iz}).$$

The rotation R_θ has $U_{R_\theta}(q(t)) = \cos\theta \mathbb{1} - \sin\theta q_x(t) q_z(t)$, so

$$\begin{aligned} R_\theta : \ (q_x, q_z) &\rightarrow (\cos\theta + \sin\theta q_x q_z)(q_x, q_z)(\cos\theta - \sin\theta q_x q_z) \\ &= ((\cos^2\theta - \sin^2\theta) q_x - 2\cos\theta \sin\theta q_z, (\cos^2\theta - \sin^2(\theta)) q_z + 2\cos\theta\sin\theta q_x) \\ &= (\cos 2\theta q_x - \sin 2\theta q_z, \cos 2\theta q_z + \sin 2\theta q_x). \end{aligned}$$

5.7.6.1. *The CNOT*

Consider a CNOT gate where the qubit c controls the target qubit t. Restricting to the subspace acted upon, the linear transformation is represented by

$$\text{CNOT} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

The functional representation is established by $U_{\text{CNOT}}(\boldsymbol{q}(0)) = \text{CNOT}$, which can be found by decomposing the above matrix.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} = \frac{1}{2}(\mathbb{1}\otimes\mathbb{1} + \sigma_z \otimes \mathbb{1})$$

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \frac{1}{2}(\mathbb{1}\otimes\sigma_x - \sigma_z \otimes \sigma_x),$$

so

$$\text{CNOT} = \frac{1}{2}(\mathbb{1} + q_{cz}(0) + q_{tx}(0) - q_{cz}(0) q_{tx}(0)).$$

The functional form of CNOT (c controls t) is hence given by

$$U_{\text{CNOT}}(\boldsymbol{q}(t)) = \frac{1}{2}(\mathbb{1} + q_{cz}(t) + q_{tx}(t) - q_{cz}(t) q_{tx}(t)).$$

The action of a CNOT is given by

$$\mathrm{CNOT}: \left\{ \begin{array}{c} (q_{cx}(t+1), q_{cz}(t+1)) \\ (q_{tx}(t+1), q_{tz}(t+1)) \end{array} \right\} = \left\{ \begin{array}{c} (q_{cx}(t)q_{tx}(t), q_{cz}(t)) \\ (q_{tx}(t), q_{cz}(t)q_{tz}(t)) \end{array} \right\}.$$

The calculation of $q_{cx}(t+1)$ can be done as follows.

$$\begin{aligned} q_{cx}(t+1) &= \frac{1}{4}(\mathbb{1} + q_{cz} + q_{tx} - q_{cz}q_{tx})\, q_{cx}\, (\mathbb{1} + q_{cz} + q_{tx} - q_{cz}q_{tx}) \\ &= \frac{1}{4}(q_{cx} + q_{cx}q_{cz} + q_{cx}q_{tx} - q_{cx}q_{cz}q_{tx} \\ &\quad +q_{cz}q_{cx} + q_{cz}q_{cx}q_{cz} + q_{cz}q_{cx}q_{tx} - q_{cz}q_{cx}q_{cz}q_{tx} \\ &\quad +q_{tx}q_{cx} + q_{tx}q_{cx}q_{cz} + q_{tx}q_{cx}q_{tx} - q_{tx}q_{cx}q_{cz}q_{tx} \\ &\quad -q_{cz}q_{tx}q_{cx} - q_{cz}q_{tx}q_{cx}q_{cz} - q_{cz}q_{tx}q_{cx}q_{tx} + q_{cz}q_{tx}q_{cx}q_{cz}q_{tx}) \\ &= \frac{1}{4}(q_{cx} + q_{cx}q_{cz} + q_{cx}q_{tx} - q_{cx}q_{cz}q_{tx} \\ &\quad -q_{cx}q_{cz} - q_{cx} - q_{cx}q_{cz}q_{tx} + q_{cx}q_{tx} \\ &\quad +q_{cx}q_{tx} + q_{cx}q_{cz}q_{tx} + q_{cx} - q_{cx}q_{cz} \\ &\quad +q_{cx}q_{cz}q_{tx} + q_{cx}q_{tx} + q_{cx}q_{cz} - q_{cx}) \\ &= q_{cx}q_{tx}, \end{aligned}$$

where, the dependency on t has been discarded.

The action of a gate on a descriptor can be found directly from the matrix representation of the gate, without the detour by its functional representation and the gymnastic of the $\mathfrak{su}(2)^{\otimes n}$ algebra. Let's exemplify the method with the case of the CNOT, which in this case consists of calculating

$$\mathrm{CNOT} \left\{ \begin{array}{c} \boldsymbol{q}_c(0) \\ \boldsymbol{q}_t(0) \end{array} \right\} \mathrm{CNOT}.$$

For the q_{cx} element, this yelds

$$\mathrm{CNOT}(\sigma_x \otimes \mathbb{1})\mathrm{CNOT} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \\
&= \sigma_x \otimes \sigma_x \\
&= q_{cx}(0) q_{tx}(0),
\end{aligned}
$$

consistently with the previous approach. But why does this work?

In fact what has been computed is

$$q_{cx}(1) = U^{\dagger}_{\text{CNOT}}(\boldsymbol{q}(0)) q_{cx}(0) U_{\text{CNOT}}(\boldsymbol{q}(0)) = q_{cx}(0) q_{tx}(0).$$

The leap to the general case, *i.e.*, to have $t+1$ and t instead of 1 and 0, follows from observing that the above equation could have been obtained by replacing $U_{\text{CNOT}}(\boldsymbol{q}(0))$ by its functional representation, use the $\mathfrak{su}(2)^{\otimes n}$ algebraic relations. But since the algebraic relations are preserved, $\boldsymbol{q}(0)$ can invariably be changed to $\boldsymbol{q}(t)$.

5.7.7. A Note to the Reader

At this stage, the reader who is curious to unravel the mystery of Bell inequality violations and of quantum teleportation is directed to §4 and §5 of the article by Deutsch and Hayden.

In fact, the explanation that the developed formalism provides to the two most famous "nonlocal" manifestations of quantum theory reaches far more than mystery breaking. It roots back quantum theory together with all other scientific theories: the act of measurement needs not to be treated as fundamentally different evolution, and it is completely local. It explores core concepts of the theory — invisible from the Shrödinger picture — that are key to good explanations. Therefore, it changes our vision of reality, making it clearer.

For the reader who is about to jump into Deutsch and Hayden's article, what follows will be useful. However, it is not needed in the present paper.

5.7.7.1. *Rotation on the Bloch Sphere*

Rotating a qubit on the Bloch sphere is described by a rotation of angle θ around the unit vector $\hat{\boldsymbol{n}}$. To distinguish this type of rotation with the rotation in the $|0\rangle$ & $|1\rangle$ plane, we denote it $\tilde{R}_{\hat{n};\theta}$.

The function representing a general rotation on the Bloch Sphere is given by

$$U_{\tilde{R}_{\hat{n};\theta}}(\boldsymbol{q}(t)) = e^{i(\theta/2)\hat{\boldsymbol{n}}\cdot\boldsymbol{q}(t)}. \tag{5.14}$$

The differences with the rotation R_θ in the $|0\rangle$ & $|1\rangle$ plane are two-fold. First, instead of exponentiating the $q_y(t)$ operator, a more general operator $\hat{\boldsymbol{n}} \cdot \boldsymbol{q}(t)$ is exponentiated. Second, the parameter becomes $\theta/2$. This is because rotating a state in the $|0\rangle$ & $|1\rangle$ plane can be seen as a rotation in the Bloch sphere with $\hat{\boldsymbol{n}} = (0, 1, 0)$, *i.e.*, fixed pointing in the y direction. However, when seen this way, a rotation of 180° on the Bloch sphere corresponds to a rotation of 90° in the plane, whence the factor of 1/2.

5.8. Appendix: Proofs of § 5.5

Proof of Theorem 4.

$$\begin{aligned}
\left(V[\![U]\!]^A\right)_{xy} &= \int dx'dy' V_{xx'}[\![U]\!]^A_{x'y'}V^\dagger_{y'y} \\
&= \int dx'dy' \langle x|V|x'\rangle \left(U^\dagger(|y'\rangle\langle x'| \otimes \mathbb{1}^{\overline{A}})U\right)\langle y'|V^\dagger|y\rangle \\
&= \int dx'dy' \left(U^\dagger(|y'\rangle\langle y'|V^\dagger|y\rangle\langle x|V|x'\rangle\langle x'| \otimes \mathbb{1}^{\overline{A}})U\right) \\
&= U^\dagger(V^\dagger|y\rangle\langle x|V \otimes \mathbb{1}^{\overline{A}})U \\
&= U^\dagger(V^\dagger \otimes \mathbb{1}^{\overline{A}})(|y\rangle\langle x| \otimes \mathbb{1}^{\overline{A}})(V \otimes \mathbb{1}^{\overline{A}})U \\
&= [\![(V \otimes \mathbb{1}^{\overline{A}})U]\!]^A.
\end{aligned}$$

$$\begin{aligned}
\left(\mathrm{tr}_B[\![U]\!]^{AB}\right)_{xy} &= \int dz [\![U]\!]^{AB}_{xz;yz} \\
&= \int dz U^\dagger(|y,z\rangle\langle x,z| \otimes \mathbb{1}^{\overline{AB}})U \\
&= \int dz U^\dagger(|y\rangle\langle x|^A \otimes |z\rangle\langle z|^B \otimes \mathbb{1}^{\overline{AB}})U \\
&= U^\dagger(|y\rangle\langle x|^A \otimes \mathbb{1}^B \otimes \mathbb{1}^{\overline{AB}})U \\
&= [\![U]\!]^A.
\end{aligned}$$

$$\begin{aligned}
\left([U]^A \odot [U]^B\right)_{x_Ax_B;y_Ay_B} &= [U]^A_{x_Ay_A}[U]^B_{x_By_B} \\
&= U^\dagger\left(|y_A\rangle\langle x_A| \otimes \mathbb{1}^B \otimes \mathbb{1}^{\overline{AB}}\right)UU^\dagger\left(\mathbb{1}^A \otimes |y_B\rangle\langle x_B| \otimes \mathbb{1}^{\overline{AB}}\right)U \\
&= U^\dagger\left(|y_A\rangle\langle x_A| \otimes |y_B\rangle\langle x_B| \otimes \mathbb{1}^{\overline{AB}}\right)U \\
&= [U]^{AB}_{x_Ax_B;y_Ay_B}.
\end{aligned}$$

□

Proof of Theorem 5.

$$
\begin{aligned}
\left(\varphi[\![U]\!]^{A}\right)_{xy} &= \mathrm{tr}\left([\![U]\!]^{A}_{xy}\rho_0\right)\\
&= \mathrm{tr}\left(U^{\dagger}(|y\rangle\langle x|\otimes \mathbb{1}^{\overline{A}})U\rho_0\right)\\
&= \mathrm{tr}\left((|y\rangle\langle x|\otimes \int dz|z\rangle\langle z|)U\rho_0U^{\dagger}\right)\\
&= \int dz\langle x,z|U\rho_0U^{\dagger}|y,z\rangle\\
&= \left(\mathrm{tr}_{\overline{A}}(U*\rho_0)\right)_{xy}.
\end{aligned}
$$

$$
\begin{aligned}
V*\varphi[\![U]\!]^{A} &= V\mathrm{tr}_{\overline{A}}(U\rho_0U^{\dagger})V^{\dagger}\\
&= \mathrm{tr}_{\overline{A}}\left((V\otimes \mathbb{1}^{\overline{A}})U\rho_0U^{\dagger}(V\otimes \mathbb{1}^{\overline{A}})^{\dagger}\right)\\
&= \mathrm{tr}_{\overline{A}}\left((V\otimes \mathbb{1}^{\overline{A}})U*\rho_0\right)\\
&= \varphi[\![(V\otimes \mathbb{1}^{\overline{A}})U]\!]^{A}\\
&= \varphi(V[\![U]\!]^{A}).
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{tr}_B\varphi\left([\![U]\!]^{AB}\right) &= \mathrm{tr}_B\left(\mathrm{tr}_{\overline{AB}}(U*\rho_0)\right)\\
&= \mathrm{tr}_{\overline{A}}(U*\rho_0)\\
&= \varphi[\![U]\!]^{A}\\
&= \varphi\mathrm{tr}_B[\![U]\!]^{AB}.
\end{aligned}
$$

□

Chapitre 6

Conclusion

(français)

> To prove more one must assume more.
>
> — Gregory Chaitin [28] —

> [T]he logic of fallibilism is that one not only seeks to correct the misconceptions of the past, but hopes in the future to find and change mistaken ideas that no one today questions or finds problematic.
>
> — David Deutsch [32] —

Les *idées* et la *rigueur* sont essentiels à de la bonne science. C'est dû à la nature faillible de la science. Par nos idées, nous proposons des théories et des explications au sujet de la réalité. Ces propositions ne seront jamais prouvées maîtresses... au contraire, elles ne peuvent qu'échouer face à une critique ou une expérience établie avec rigueur.

J'ai l'impression que les scientifiques sont souvent si attachés à leur fonds de recherche, à leur carrière et à leur propres idées qu'ils refusent de voir l'échec d'une de leur proposition comme un progrès. La position naturelle devient donc celle de l'expert de la rigueur et de la critique, très froid aux idées extraordinaires.

Pour ne pas freiner la créativité, je fais voeu de toujours trouver belle l'idée d'échouer et de changer d'avis en science. Ainsi je profite de cette dernière section de ma thèse pour libérer mes idées restantes, bien que spéculatives et sans rigueur.

6.1. Un bref retour sur le hasard accidentel.

Je souhaite revenir sur un élément des conclusions du chapitre 4. Le hasard accidentel («*incidental randomness*») dans l'univers, soit l'information algorithmique irréductible indépendante de Ω, pourrait en fait être signe de multiplicité dans un univers pourtant simple. À l'instar des modèles algorithmiques, un ensemble d'objets (même infini) peut être beaucoup plus simple qu'un seul d'entre eux. Par exemple, $\mathbb{N}$ est plus élégant que

$$98153791587745457509645604951640958459924890869086200\ldots5.$$

Ainsi, un multivers algorithmiquement simple est tout à fait compatible avec une tranche apparemment complexe. C'est en particulier le cas d'un univers unitaire simple qui, à la lumière d'Everett, ne contredit pas la complexité croissante de mesures de systèmes quantiques. Un argument semblable pourrait également tenir compte de la complexité des nombreuses constantes physiques dans un multivers cosmologique pourtant plus simple.

6.2. Ω dans le monde naturel

À la lumière du lien entre la sophistication et la profondeur logique, il s'avère que notre proposition de l'émergence impliquerait que les structures émergentes autour de nous partagent de l'information mutuelle avec le problème d'arrêt. Celle-ci peut être cristallisée dans un préfixe de Ω. En supposant que «nous» qualifions de structures émergentes et que notre ADN encode suffisamment de nos caractéristiques pertinentes, alors de l'information à propos de Ω devrait être inscrite de manière holographique dans notre ADN, c'est-à-dire, par la profondeur de sa cause la plus plausible. Dans la mesure où les bits de Ω devraient aussi être inscrits dans toute autre structure émergente, nous leur serions alors algorithmiquement connectés. Cette conséquence nous était imprévisible lorsque Geoffroy et moi avons initié notre travail sur l'émergence. Mais on ne choisit pas ce que nos idées impliquent, et des conséquences contre-intuitives ne devraient jamais être un critère pour les exclure. Au lieu, voyons cela plus en profondeur.

Ω *est le nombre incompressible le plus naturel.* On peut en parler, on peut le définir mathématiquement, et il a la propriété remarquable que pour tout k, *presque tous* les nombres naturels partagent avec lui plus de k bits d'information mutuelle. En effet, seul un nombre fini de nombres naturels est inférieur à $\boldsymbol{B}(k)$ et tous les autres peuvent servir à calculer les k premiers bits de Ω grâce à leur taille gigantesque qui leur permet de «stabiliser k bits sur l'horloge Ω» (voir chapitre 4). Cette propriété souligne que toute pièce d'information capable de causer une machine universelle de nommer un grand nombre — ou d'exécuter beaucoup d'étapes de calcul avant de s'arrêter — partage de l'information avec Ω. Alors Ω *est peut-être notre meilleur moyen de représenter l'infini*: plus le nombre que l'on peut nommer est grand, plus on se rapproche de Ω et *vice versa*.

Ω *est rempli de vérités mathématiques utiles* [18]. En effet, le statut d'arrêt de certains courts programmes permet de déterminer si de nombreuses conjectures mathématiques sont vraies ou non. Par exemple, connaître le statut d'arrêt du programme qui s'arrête dès qu'il trouve un nombre pair supérieur à 2 qui n'est pas la somme de deux nombres premiers, c'est connaître la vérité de la *conjecture de Goldbach*. Mais Ω *est aussi rempli de vérités «cosmiques»*, car ultimement, cela demeure une propriété de l'univers de savoir si la motion d'un engin d'arrête ou non... et Ω encode joliment ces propriétés. Ainsi, du moins dans la limite du temps qui passe, les vérités mathématiques et cosmiques ne seraient donc pas si différentes... toutes deux réunies dans Ω.

La même surprise en métabiologie

Ω surgit aussi en *metabiologie* [27, 28]. Et là aussi, cela survient à la surprise de l'auteur de la théorie, soit Gregory Chaitin même. La métabiologie est un modèle jouet de la biologie qui cherche à mieux comprendre comment de *nouveaux* gènes peuvent survenir et ainsi expliquer la *créativité* inhérente à la biosphère. Dans le modèle, il n'y a pas d'environnement, pas de compétition, pas de sexe; seulement des organismes qui sont représentés par des programmes. Comme Chaitin l'a observé, l'humanité devait développer ses propres langages de programmation pour se rendre compte que le logiciel («*software*») était déjà présent dans le monde entier: *La vie est pleine de logiciels* — exécutés dans le langage naturel de l'ADN. “Software is the reason for the plasticity of the biosphere — normal machines are rigid, mechanical, dead. Software is alive”! La métabiologie est donc une théorie mathématique de la vie en tant que logiciel évolutif. Les organismes cherchent à optimiser une tâche d'adaptation («*fitness*») qui demande une créativité mathématique illimitée : il s'agit de nommer un grand nombre. Des mutations aléatoires se produisent (dans l'espace des programmes) et les nouveaux organismes (nouveaux programmes) sont conservés s'ils peuvent nommer un plus grand nombre que leur prédécesseur. Pour éviter que l'évolution ne soit prise dans une boucle, l'accès à un oracle du problème d'arrêt est fourni. Les organismes les mieux adaptés ont donc de plus en plus de connaissances sur Ω, encore une fois, par leur capacité à nommer un grand nombre.

La créativité des organismes survient puisqu'ils sont continuellement défiés par une question mathématique ouverte. Dans le modèle de Chaitin, la porte de cette créativité est l'accès à l'oracle. Mais cela peut-il être modifié pour plutôt permettre à l'organisme d'interagir avec un environnement *émergent* rempli d'information d'arrêt? Cela pourrait commencer à paraître réaliste... Mais le problème semble seulement repoussé à un niveau supérieur, puisqu'il faut maintenant expliquer comment cet environnement (constitué en partie d'autres organismes) a recueilli cette information d'arrêt. Bien que notre proposition d'émergence suggère que l'environnement contienne de l'information d'arrêt, elle ne modélise pas comment il l'a obtenue. Toutefois, l'auto génération de l'information d'arrêt dans un univers infini n'est peut-être pas si difficile à expliquer : un programme qui ajoute 1 à un compteur imprimé fait étoffe de modèle-jouet.

Nos mathématiques dans le monde naturel

J'ai toujours pensé que la physique était mathématique. Je crois toujours que c'est le cas, mais je préfère penser que nos mathématiques sont physiques, car cette idée a plus de portée. Par «nos mathématiques», j'entends ici le type de mathématiques enseigné dans les départements de maths, et non pas quelque système formel avec quelque logique que ce soit. J'appellerai ce domaine beaucoup plus vaste «les Mathématiques». Mais comment *nos m*athématiques ont-elles été sélectionnées, parmi les nombreuses possibilités offertes par *les M*athématiques?

Ferdinand Gonseth a été cité *"La logique est d'abord une science naturelle"* [48], et je ne pourrais être plus en accord. Il n'existe pas de mathématicien vierge. Il a de l'ADN qui porte l'histoire de ses ancêtres. Il a des perceptions, des émotions, des pensées... Plus important encore, il vit dans cet univers, soumis à ses lois, qui, que cela lui plaise ou non, contraignent le fonctionnement de son cerveau, de son crayon et de son papier. Nous ne devrions donc pas nous surprendre qu'il trouve beau quelque chose comme $e^{i\pi}+1=0$,

$\sum \frac{1}{n^2} = \frac{\pi^2}{6}$, ou l'un des théorèmes les plus abstraits qu'il utilise dans son travail. Il les trouve plus beaux que des théorèmes quelconques de systèmes formels quelconques des Mathématiques. Ces derniers ne signifient rien pour l'Homme, et même peut-être qu'ils ne signifient rien pour l'Univers. Mais ceux qu'ils trouvent beaux et ceux qu'il utilise sont construits à partir d'idées et d'abstractions qui descendent de notre monde physique, inévitablement!

Par exemple, lorsque Galois découvre un chapitre inconnu des mathématiques la nuit qui précède sa mort, il découvre *ses* mathématiques, qui sont son expression créatrice, sa vision du monde — du moins de son monde abstrait intérieur — qui à la lumière des idées abstraites déjà en place, n'a que de raisons de ressembler à notre monde extérieur. Une fois ses écrits compris, ses mathématiques deviennent *les nôtres*. Cette nouvelle branche des mathématiques n'implique aucune modification aux axiomes de la théorie des ensembles, puisqu'elle est construite à partir d'objets préalablement bien définis. Mais cela n'est pas le cas de l'hypothèse du continu, par exemple, qui est indépendante des axiomes actuels de la théorie des ensembles. Des factions de mathématiciens peuvent être créées, celle pour et celle contre l'hypothèse, débattant de la façon dont nous devrions élargir nos mathématiques à partir des possibilités infinies des Mathématiques. Mais qui a raison? Comment devrions-nous décider? Eh bien... nous semblons avoir voté contre le cinquième postulat d'Euclide, mais pour l'axiome de choix. Et ces choix semblent être basés sur l'utilité, par exemple, pour traiter des espaces courbes ou pour enrichir la théorie de la mesure.

Tout ce processus ressemble à ce que font les scientifiques. Proposer des idées, celles qui sont utiles sont conservées jusqu'à ce qu'elles se révèlent contradictoires ou non pertinentes à la lumière d'une idée plus ambitieuse. En un mot, *l'incomplétude amène le faillibilisme en mathématiques*.

Supernova

Le carbone, l'oxygène, l'azote, le fer, et tous les autres éléments chimiques qui fabriquent la vie autour de nous — et en nous — ont jadis été transmutés dans les étoiles. Cette matière robuste et féconde a traversé l'espace et le temps avant de former cette Terre, la garnissant d'une myriade d'interactions et de possibilités qui ont conduit jusqu'à nous. *Est-ce que l'information du problème d'arrêt, robuste et féconde, pourrait être générée dans l'univers et ultimement infiltrer nos mathématiques?*

Certains objecteront peut-être [55, 57] que l'inégalité du traitement algorithmique des données empêche nos mathématiques de trop en savoir sur l'univers, car aucun processus simple calculable ne peut augmenter de manière significative l'information mutuelle entre deux objets... Mais cela ne vaut que s'ils sont traités indépendamment, à savoir, sans interaction. Nous devons toutefois prendre en compte dans quelle mesure nos mathématiques, à travers nous, *font partie de l'univers*.

Gödel a dit [47] :

> *Namely, it turns out that in the systematic establishment of the axioms of mathematics, new axioms, which do not follow by formal logic from those previously established, again and again become evident. It is not at all excluded by the negative results mentioned earlier that nevertheless every clearly posed mathematical yes-or-no question is solvable in this way. For it is just this becoming evident of more and more new axioms on the basis of the meaning of the primitive notions that a machine can not imitate.*

Ramanujan a écrit

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{n=0}^{\infty} \frac{(4n)!}{(n!)^4} \times \frac{1103 + 26390n}{(4 \times 99)^{4n}}.$$

Sa formule est vraie, mais n'a été prouvée que 70 ans après son écriture! Il a aussi dit : "An equation has no meaning to me unless it expresses a thought of God". Comment cela peut-il être expliqué scientifiquement?

Bien entendu, derrière tout cela, il y a une explication scientifique sans intervention divine ni distinction fondamentale entre l'homme et la machine. Cette explication rendra peut-être compte de ce flot de vérités mathématiques, de vérités cosmiques, de morceaux de Ω qui s'infiltrent dans nos mathématiques. Puisque l'information est encore plus fongible que les éléments chimiques, qu'est-ce qui pourrait limiter cela?

Et nous?

David Deutsch suggère [32, 34] que la notion de *connaissance* devrait être davantage considérée en physique. Les processus que peuvent subir la matière, les champs et l'espace-temps dépendent fondamentalement de la connaissance avoisinante. De l'air, de l'eau et de la lumière peuvent être changés en arbres; des régions à la surface de la

Terre peuvent être plus froides et plus sombres que les régions les plus profondes de l'espace intergalactique; des astéroïdes peuvent avoir leur trajectoire déviée... Tout cela et bien plus grâce à la connaissance appropriée!

À la lumière du principe physique de Church-Turing, l'information d'arrêt permet de calculer les trajectoires possibles d'un engin de calcul universel, et donc les processus physiques possibles. De ce fait, la croissance des connaissances de l'humanité pourrait trouver plus qu'une métaphore dans le flot de vérités cosmiques que nous recueillions scientifiquement et mathématiquement.

Mais nous, humains, ne serions pas que les spectateurs de ce flot. Nous contribuons. L'ADN, les systèmes nerveux, le langage, l'écriture, le rejet de l'autorité comme assise du savoir, les ordinateurs, les ordinateurs quantiques sont des révolutions successives du traitement de l'information et de l'érection de la connaissance dans lequel nous avons joué un rôle central. Si l'univers auto génère des bits de Ω de façon bien plus élaborée qu'en ajoutant 1 à un compteur, alors nous en faisons certainement partie. Et ça vient probablement juste de commencer!

Chapitre 7

Conclusion

(English)

> To prove more one must assume more.
>
> — Gregory Chaitin [28] —

> [T]he logic of fallibilism is that one not only seeks to correct the misconceptions of the past, but hopes in the future to find and change mistaken ideas that no one today questions or finds problematic.
>
> — David Deutsch [32] —

Ideas and *rigour* are essential to good science. This due to the fallible nature of science. Through our ideas, we propose theories and explanations about reality. These propositions will never be proved right... on the contrary, they can only fail in the face of a criticism or an experiment established with rigour.

I have the impression that scientists are often so bound to their research fund, their career and their own ideas that they refuse to see the failure of one of their proposals as progress. The natural position becomes that of the expert of rigour and criticism, very cold with extraordinary ideas.

In order not to curb creativity, I vow to always find beauty in failing and in changing my mind in science. So I take this last section of my thesis to free my remaining ideas, even if they are speculative and without rigour.

7.1. A Short Retake on Incidental Randomness

Let me come back on an element of conclusion of Chapter 4. Incidental randomness in the Universe, the one that is independent of Ω, may simply be a sign of multiplicity in the Universe. As for algorithmic models, many objects — even infinitely many — may be simpler than a single one of them. For instance, $\mathbb{N}$ is nicer than

$$981537915877454575096456049516409584599248908690862 00\ldots 5.$$

So, a simple Many-Universe (multiverse) is compatible with an apparently complex slice of Universe. Everettly-interpreted, this accounts for the unbounded complexity of increasing series of measurement in a nonetheless simple uniatry Universe. Similar arguments could account for the apparent randomness in the many physical constants of the Universe within a simpler cosmological multiverse.

7.2. On the Naturalness of Ω

In the light of the connexion between sophistication and busy beaver depth, it turns out that our proposal of emergence conjectures that emergent structures around us share mutual information with the halting problem that can be crystallized in a prefix of Ω. Assuming that "we" qualify as emergent structures, and that our DNA encodes enough of our relevant features, then some information about Ω should be holographically written in our DNA, *i.e.*, through the depth of its most plausible computational cause. Inasmuch as bits of Ω should similarly be written in any other emergent structure, we should then be algorithmically connected to them. This consequence was unforeseeable when Geoffroy and I initiated the work on emergence. But we cannot choose what our ideas imply, and counter-intuitive consequences should never be a criterion to rule them out. Instead, let's investigate this further.

Ω *is the most natural incompressible sequence.* For we can speak of it, we can define it mathematically precisely, and it has the remarkable property that for any k, *almost all* natural numbers share with it more than k bits of mutual information. This is because only finitely many natural numbers are smaller than $\boldsymbol{B}(k)$ and all the others algorithmically know the first k bits of Ω from their overwhelming size that endow them with the ability to "stabilize k bits on the Ω clock" (*cf.* Chapter 4). This property underlines that any piece of information that can cause a universal computer to name a large number — or to run for many steps before halting — shares information with Ω. So Ω *is perhaps our best way to represent infinity*: The bigger the number one can name, the closer one gets to Ω, and *vice versa*.

Ω *is stuffed with useful mathematical truths* [18]. Indeed, the halting status of some short programs can determine whether many mathematical conjectures are true or not. For instance, knowing the halting status of the program that halts as soon as it finds an even number greater than 2 that is not the sum of two primes is knowing the truth of *Goldbach's conjecture*. But Ω *is also stuffed with "cosmic" truths*, because ultimately, it is a property of the Universe whether a motion of some device comes to a halt or not, and Ω nicely encodes those halting properties. Hence, at least in the limit as time goes on, mathematical and cosmic truths shall not be so different — both encompassed in Ω.

The Same Surprise in Metabiology

Ω also crops out in *metabiology* [27, 28]. There too, it happens to the surprise of the theory's author, *i.e.*, Gregory Chaitin himself. Metabiology is a toy model of biology that seeks to better understand how *new* genes can arise and explain the inherent *creativity* of the biosphere. In the model there is no environment, no competition, no sex; Only organisms, which are represented by programs. As Chaitin observed, humanity needed to develop his own artificial programming languages to realize that software was already all over the globe: *Life is full of software* — executed in the natural language of DNA. "Software is the reason for the plasticity of the biosphere — normal machines are rigid, mechanical, dead. Software is alive"! So metabiology is a mathematical theory of *life as evolving software*. Organisms seek to optimize a fitness task that asks for unlimited mathematical creativity: Naming a large number. Random mutations occur (in program space), and new organisms (new programs) are kept if they can name a larger number than their predecessor. To prevent evolution from being stuck in a loop, access to an oracle of the halting problem is provided. So fitter and fitter organisms have more and more knowledge about Ω, again, by their ability to name a large number.

Creativity is forced on the organisms by constantly challenging them with an open-ended mathematical question. In Chaitin's model, the door to creativity is the access to the oracle. Can this be relaxed by instead allowing the organism to interact with an *emergent* environment that is deep enough to carry halting information? This may actually start to look realistic... But the problem may just be lifted one level up, since one now needs to explain how this environment (in part made of other organisms) collected that halting information. Although our emergence proposal suggests that the environment does contain halting knowledge, it does not model how it got it. But self-generation of halting knowledge in an endless Universe may not be so hard to explain, as it could be toy-modelled by a program ever adding one to a printed count.

On the Naturalness of our *m*athematics

I used to think that physics was mathematical. This is still true, I think, but I'd rather regard our mathematics as being physical, since I think that this idea reaches further. By "our mathematics", I mean here the kind of mathematics that is taught in a math department, and not any sort of formal system with any sort of logic... Let me call this much bigger realm "Mathematics". How have our mathematics been selected, out of the many possibilities offered by Mathematics?

"La logique est d'abord une science naturelle" [48]. This quote from Ferdinand Gonseth states that logic is, first of all, a natural science, and I could not agree more. There is no such thing as a blank state mathematician. He has DNA, perceptions, emotions, memories, thoughts... More importantly, he lives in this Universe, subjected to its laws, which, whether he likes it or not, constrain how its brain, pen and paper operate for him to do his work. And the same holds for all of its ancestors, mathematicians or not. Henceforth, we shall not be surprised that he finds beautiful something like $e^{i\pi} + 1 = 0$, $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$, or any of the most abstract theorems he uses in his work. He finds them more beautiful than some random theorems of some random formal axiomatic system of Mathematics. The latter mean nothing for humans, and even, maybe they mean nothing for the Universe. But the former notions are built from ideas and abstractions that descend from our physical world, inevitably!

For instance, when Galois comes up with an unknown chapter of mathematics the night before he dies, he comes up with *his* mathematics, which is his creative expression, his vision about the world — at least about his own abstract inner world — which has only reasons to resemble our outer concrete world. Once his writings were understood, his mathematics became *ours*. This new branch of mathematics involved no modification to the axioms of set theory, since it was built at a high level on previously well-defined objects. But this isn't the case of the continuum hypothesis, for example, which is independent from the current axioms of set theory. Factions of mathematicians may be created, those for and those against the hypothesis, debating on how we should enlarge our mathematics from the limitless Mathematics. Which faction is right? How should we decide? Well... we seem to have voted against Euclid's fifth postulate but for the axiom of choice. And those choices were based on utility, *e.g.*, to address curved spaces or to enrich measure theory.

This whole process resembles what scientists do. Propose ideas, those that are useful should be kept, until they are proven contradictory or irrelevant in the light of a more outreaching idea. In a nutshell, *incompleteness takes fallibilism into mathematics*.

Supernova

Carbon, Oxygen, Nitrogen, Iron, and all the other chemical elements that build life around us — and in us — have once been transmuted in stars. This matter, robust and fecund, made its way across space and time before forming this Earth, providing it with a myriad of interactions and opportunities to now build us. *Can halting information, robust and fecund, be generated in the Universe and ultimately reaches our mathematics?*

Some may object [**55**, **57**] that the algorithmic data processing inequality prevents our mathematics from knowing too much about the Universe, since no simple computable process can significantly increase the mutual information between two objects... But this only holds if they are independently processed, namely, non-interacting; We must take into account to which extent our mathematics, through us, are *a part of the Universe*.

Gödel said [**47**]:

> *Namely, it turns out that in the systematic establishment of the axioms of mathematics, new axioms, which do not follow by formal logic from those previously established, again and again become evident. It is not at all excluded by the negative results mentioned earlier that nevertheless every clearly posed mathematical yes-or-no question is solvable in this way. For it is just this becoming evident of more and more new axioms on the basis of the meaning of the primitive notions that a machine can not imitate.*

Ramanujan came up with

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{n=0}^{\infty} \frac{(4n)!}{(n!)^4} \times \frac{1103 + 26390n}{(4 \times 99)^{4n}}.$$

His formula is true, but has only been proven 70 years after he wrote it! He claimed that "[a]n equation has no meaning to me unless it expresses a thought of God". How can this be explained scientifically?

Of course, behind it all there is a scientific explanation with no divine intervention and no fundamental distinction between man and machine. This explanation would perhaps account for the aforementioned flow of mathematical truth, cosmic truth, pieces of Ω into our mathematics. Since information is even more fungible than chemical elements, what could limit this?

And Us?

David Deutsch suggests [32, 34] that *knowledge* considerations should be incorporated to physics. The processes that matter, fields and spacetime may undergo are fundamentally dependent upon knowledge. Air, water and light can be changed into trees, regions at the surface of the Earth can get colder and darker than the deepest regions of intergalactic space, asteroids can have their trajectories deviated... All this, and much more, with the appropriate knowledge!

In the light of the physical Church-Turing's principle, halting knowledge permits to compute the possible motions of a Universal computer, so the possible physical processes. Thereby, humanity's growth of knowledge may find more than a metaphor in the inwards flow of cosmic truths that we scientifically and mathematically gather.

But we humans are not merely spectators of this flow. We contribute. DNA, nervous systems, language, writing, rejection of authority as knowledge providers, computers and quantum computers are successive revolutions of information processing in which we play a central role. If the Universe self-generates bits of Ω in a much fancier way than adding 1 to a counter, then we are a part of this. And it has probably just begun!

Bibliography

[1] Retrieved from https://en.wikipedia.org/wiki/Emergence, April 2019.

[2] Philip W Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[3] Luís Antunes, Bruno Bauwens, André Souto, and Andreia Teixeira. Sophistication vs logical depth. *Theory of Computing Systems*, 60(2):280–298, 2017.

[4] Luís Antunes and Lance Fortnow. Sophistication revisited. *Theory of Computing Systems*, 45(1):150–161, 2009.

[5] Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental realization of Einstein-Podolsky-Rosen-Bohm Gedankenexperiment: a new violation of Bell's inequalities. *Physical Review Letters*, 49(2):91, 1982.

[6] Nihat Ay, Markus Muller, and Arleta Szkola. Effective complexity and its relation to logical depth. *IEEE transactions on information theory*, 56(9):4593–4607, 2010.

[7] Leslie E Ballentine. *Quantum mechanics: a modern development*. World Scientific Publishing Company, 1998.

[8] Ämin Baumeler and Stefan Wolf. Causality–Complexity–Consistency: Can space-time be based on logic and computation? In *Time in Physics*, pages 69–101. Springer, 2017.

[9] Bruno Bauwens. On the equivalence between minimal sufficient statistics, minimal typical models and initial segments of the halting sequence. *arXiv preprint arXiv:0911.4521*, 2009.

[10] Bruno Bauwens. *Computability in statistical hypotheses testing, and characterizations of independence and directed influences in time series using Kolmogorov complexity*. PhD thesis, Ghent University, 2010.

[11] Charles Alexandre Bédard and Geoffroy Bergeron. An algorithmic approach to quantify emergence. In *Conférence de l'Institut transdisciplinaire d'information quantique*, 2018.

[12] Mark A Bedau and Paul Ed Humphreys. *Emergence: Contemporary readings in philosophy and science.* MIT press, 2008.

[13] John S Bell. On the Einstein Podolsky Rosen paradox. *Physics*, 1(3):195–200, 1964.

[14] John S Bell. *Speakable and unspeakable in quantum mechanics: Collected papers on quantum philosophy*. Cambridge University Pressress, 2004.

[15] Charles H Bennett. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982.

[16] Charles H Bennett. Logical depth and physical complexity. *The Universal Turing Machine A Half-Century Survey*, pages 227–257, 1988.

[17] Charles H Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters*, 70(13):1895, 1993.

[18] Charles H Bennett and Martin Gardner. The random number omega bids fair to hold the mysteries of the universe. *Scientific American*, 241(5):20–34, 1979.

[19] Gilles Brassard and Paul Raymond-Robichaud. The equivalence of local-realistic and no-signalling theories. *arXiv preprint arXiv:1710.01380*, 2017.

[20] Gilles Brassard and Paul Raymond-Robichaud. Parallel lives: A local-realistic interpretation of "nonlocal" boxes. *Entropy*, 21(1):87, 2019.

[21] Herbert B Callen and Theodore A Welton. Irreversibility and generalized noise. *Physical Review*, 83(1):34, 1951.

[22] Jean-Sébastien Caux and Jorn Mossel. Remarks on the notion of quantum integrability. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02023, 2011.

[23] Gregory J Chaitin. On the length of programs for computing finite binary sequences. *Journal of the ACM*, 13(4):547–569, 1966.

[24] Gregory J Chaitin. A Theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340, 1975.

[25] Gregory J Chaitin. *Meta maths!: the quest for omega*. Vintage, 2006.

[26] Gregory J Chaitin. The halting probability omega: Irreducible complexity in pure mathematics. *Milan Journal of Mathematics*, 75(1):291–304, 2007.

[27] Gregory J Chaitin. To a mathematical theory of evolution and biological creativity. Technical report, Department of Computer Science, The University of Auckland, New Zealand, 2010.

[28] Gregory J Chaitin. *Proving Darwin: making biology mathematical*. Vintage, 2012.

[29] Daniel C Dennett. Real patterns. *The journal of Philosophy*, 88(1):27–51, 1991.

[30] David Deutsch. Quantum theory, the Church–Turing principle and the universal quantum computer. *Proceedings of the Royal Society A. Mathematical, Physical and Engineering Sciences*, 400(1818):97–117, 1985.

[31] David Deutsch. Quantum computational networks. *Proceedings of the Royal Society A. Mathematical, Physical and Engineering Sciences*, 425(1868):73–90, 1989.

[32] David Deutsch. *The beginning of infinity: Explanations that transform the world*. Penguin UK, 2011.

[33] David Deutsch. Vindication of quantum locality. *Proceedings of the Royal Society A. Mathematical, Physical and Engineering Sciences*, 468(2138):531–544, 2011.

[34] David Deutsch. Constructor theory. *Synthese*, 190(18):4331–4359, 2013.

[35] David Deutsch and Patrick Hayden. Information flow in entangled quantum systems. *Proceedings of the Royal Society A. Mathematical, Physical and Engineering Sciences*, 456(1999):1759–1774, 2000.

[36] Bryce S DeWitt and Neill Graham. *The many worlds interpretation of quantum mechanics*. Princeton University Press, 1973.

[37] Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47(10):777, 1935.

[38] Hugh Everett III. The theory of the universal wave function. 1956.

[39] Hugh Everett III. "Relative state" formulation of quantum mechanics. *Reviews of modern physics*, 29(3):454, 1957.

[40] Richard P Feynman. Simulating physics with computers. *International Journal of Theoretical Physics*, 21(6):467–488, 1982.

[41] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222(594-604):309–368, 1922.

[42] Peter Gács. On the symmetry of algorithmic information. In *Doklady Akademii Nauk*, volume 218, pages 1265–1267. Russian Academy of Sciences, 1974.

[43] Péter Gács, John T Tromp, and Paul MB Vitányi. Algorithmic statistics. *IEEE Transactions on Information Theory*, 47(6):2443–2463, 2001.

[44] Murray Gell-Mann and Seth Lloyd. Information measures, effective complexity, and total information. *Complexity*, 2(1):44–52, 1996.

[45] Murray Gell-Mann and Seth Lloyd. Effective complexity. In *Murray Gell-Mann: Selected Papers*, pages 391–402. World Scientific, 2010.

[46] Kurt Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für mathematik und physik*, 38(1):173–198, 1931.

[47] Kurt Gödel. The modern development of the foundations of mathematics in the light of philosophy. In Oxford University Press, editor, *Collected Works*, volume 3, 1961.

[48] Ferdinand Gonseth. *Les mathématiques et la réalité. Essai sur la méthode axiomatique*. 1936.

[49] Daniel Gottesman. The Heisenberg representation of quantum computers. *arXiv preprint quant-ph/9807006*, 1998.

[50] Clare Hewitt-Horsman and Vlatko Vedral. Developing the Deutsch–Hayden approach to quantum mechanics. *New Journal of Physics*, 9(5):135, 2007.

[51] Andreï N Kolmogorov. Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11, 1965.

[52] Andreï N Kolmogorov. Talk at the Information Theory Symposium in Tallinn. *Estonia (then USSR)*, 1974.

[53] Moshe Koppel. Complexity, depth, and sophistication. *Complex Systems*, 1(6):1087–1091, 1987.

[54] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–708. Springer, 2013.

[55] Leonid A Levin. Laws of information conservation (nongrowth) and aspects of the foundation of probability theory. *Problemy Peredachi Informatsii*, 10(3):30–35, 1974.

[56] Leonid A Levin. Private communication. e-mails to P. Vitányi, Feb. 2002.

[57] Leonid A Levin. Forbidden information. *Journal of the ACM*, 60(2):9, 2013.

[58] George H Lewes. *Problems of life and mind*. 1875.

[59] Ming Li and Paul Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, New York, 2008.

[60] Luc Longpré. *Resource bounded Kolmogorov complexity, a link between computational complexity and information theory*. PhD thesis, 1986.

[61] James Clerk Maxwell. *Theory of heat*. Cambridge University Press, 1871.

[62] Alexey Milovanov. Some properties of antistochastic strings. *Theory of Computing Systems*, 61(2):521–535, 2017.

[63] An A Muchnik and Andrei E Romashchenko. Stability of properties of Kolmogorov complexity under relativization. *Problems of information transmission*, 46(1):38–61, 2010.

[64] Timothy O'Connor and Hong Yu Wong. Edward N. Zalta, ed.«Emergent Properties», 2012.

[65] Paul Raymond-Robichaud. *L'équivalence entre le local-réalisme et le principe de non-signalement*. PhD thesis, Université de Montréal, https://papyrus.bib.umontreal.ca/xmlui/handle/1866/20497, 2018.

[66] Roger D Rosenkrantz. *ET Jaynes: Papers on probability, statistics and statistical physics*, volume 158. Springer Science & Business Media, 2012.

[67] William D Ross. Aristotle's metaphysics. A revised text with introduction and commentary. 1925.

[68] Gian-Carlo Rota. *Indiscrete thoughts*. Birkhäuser, 1997.

[69] Paul A Schilpp. *Albert Einstein: philosopher-scientist*, volume 7. Open Court Publishing Co., 3d revised edition, 1970.

[70] Laurent Schwartz and Institut de mathématique (Strasbourg). *Théorie des distributions*, volume 2. Hermann Paris, 1957.

[71] Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[72] Alexander Shen, Vladimir A Uspensky, and Nikolay Vereshchagin. *Kolmogorov complexity and algorithmic randomness.* MCCME (Russian), 2013. English translation: http://www.lirmm.fr/˜ashen/kolmbook-eng.pdf.

[73] Ray J Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22, 1964.

[74] Alan M Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London mathematical society*, 2(1):230–265, 1937.

[75] Nikolai Vereshchagin and Paul Vitányi. Kolmogorov's structure functions with an application to the foundations of model selection. In *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, pages 751–760. IEEE, 2002.

[76] Nikolay Vereshchagin and Alexander Shen. Algorithmic statistics: forty years later. In *Computability and Complexity*, pages 669–737. Springer, 2017.

[77] Paul M Vitányi. Meaningful information. *IEEE Transactions on Information Theory*, 52(10):4617–4626, 2006.

[78] David Wallace. Everett and structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 34(1):87–105, 2003.

[79] David Wallace. *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford University Press, 2012.

[80] David Wallace and Christopher G Timpson. Non-locality and gauge freedom in Deutsch and Hayden's formulation of quantum mechanics. *Foundations of Physics*, 37(7):1069–1073, 2007.

[81] Wojciech H Zurek. Algorithmic randomness and physical entropy. *Physical Review A*, 40(8):4731, 1989.